Published in final edited form as:
J Dent Oral Craniofac Epidemiol. 2013 October 1; 1(3): 19–33.

Improving Patient Care Using the Johnson-Neyman Analysis of Heterogeneity of Treatment Effects According to Individuals' Baseline Characteristics

Ann A. Lazar[1], **Stuart A. Gansky**[2], **Donald D. Halstead**[3], **Anthony Slajs**[4], and **Jane A. Weintraub**[5]

[1]The University of California, San Francisco (UCSF) School of Dentistry, Division of Oral Epidemiology and Dental Public Health, & School of Medicine, Division of Biostatistics, USA

[2]The University of California, San Francisco (UCSF) School of Dentistry, Division of Oral Epidemiology and Dental Public Health, USA

[3]Harvard School of Public Health, USA

[4]The University of California, San Francisco (UCSF) School of Dentistry, Division of Oral Epidemiology and Dental Public Health, USA

[5]The University of North Carolina at Chapel Hill School of Dentistry, USA

Abstract

Objective—Because each patient's baseline (pre-treatment) characteristics differ (e.g., age, sex, socioeconomic status, ethnicity/race, biomarkers), treatments do not work the same for every patient-some can even cause detrimental effects. To improve patient care, it is critical to identify such heterogeneity of treatment effects. But the standard analytic approach dichotomizes baseline characteristics (low vs. high) which often leads to a loss of critical patient-care information and power to detect heterogeneity, as the results may depend strongly on the cut-points chosen. A more powerful analytic approach is to analyze baseline characteristics (i.e., covariates) measured on a continuous scale that retains all of the information available for the covariate.

Methods—In this article, we show how the Johnson-Neyman (J-N) method can be used to identify the prognostic and predictive value of baseline covariates measured on a continuous scale - findings that often cannot be determined using the traditional dichotomized approach. As an example, we used the J-N method to explore treatment effects for varying levels of the biomarker salivary mutans streptococci (MS) in a randomized clinical prevention trial comparing fluoride varnish with no fluoride varnish for 376 initially caries-free high-risk children, all of whom received oral health counseling.

Results—The J-N analysis showed that children with higher baseline MS values who were randomized to receive fluoride varnish had the poorest dental caries prognosis and may have benefitted most from the preventive agent.

Corresponding Author: Ann A. Lazar. ann.lazar@ucsf.edu.

Conclusion—Such methods are likely to be an important tool in the field of personalized oral health care.

Keywords

Fluoride varnish; interaction; Johnson-Neyman method; salivary mutans streptococci; subgroup analysis; treatment-effect heterogeneity; personalized medicine; precision medicine

To assess preventive and therapeutic benefits, modern dental research relies on clinical trials that are extremely expensive and time consuming. It is therefore critical that the greatest amount of information be extracted from them. At the same time, it is well known that the treatments evaluated in these trials, including those that are preventive and curative, do not work the same for every patient and that some may experience detrimental effects (1–6). To identify who may or may not benefit from a particular treatment, investigators frequently use subgroup analyses to evaluate 'treatment-effect heterogeneity' according to patients' baseline characteristics.

We define "subgroup analysis" here as any treatment effect evaluation for a specific end point in subgroups of study participants, while subgroups are constructed according to participants' baseline characteristics (e.g., socioeconomic status, age, sex, race/ethnicity, biomarkers). The end point is usually the primary endpoint, such as a measure of treatment efficacy or safety. The treatment effect, a comparison between the treatment groups, is either measured on the absolute scale (e.g., arithmetic difference) or relative scale (e.g., relative risk or odds ratio). Subgroup analyses are often listed as a primary or secondary study objective, and the research question usually posed may depend on how the baseline characteristic is measured: continuous versus dichotomized. Baseline characteristics measured on a continuous scale are often categorized into two (dichotomized) or more groups (e.g., salivary mutans streptococci (MS) level dichotomized as low vs. high), resulting in a loss of information and power (7), while the results from these dichotomized analyses may depend on the chosen cut-point. However, for baseline characteristics measured on a continuous scale (e.g., age, baseline MS), an investigator can ask how treatment effects vary by the levels of those characteristics.

Identifying treatment-effect heterogeneity provides critical patient information needed to tailor treatments according to an individual's characteristics and to ultimately lead to more effective oral health care delivery for personalized medicine (6). The purpose of this article is to show how baseline characteristics measured on a continuous scale can be analyzed using the Johnson-Neyman (J-N) method (8–18). This method has been applied to many different types of studies in education, psychology and medicine (19–22), but to our knowledge has not been applied to oral health. We illustrate this J-N method with an analysis of the Center to Address Disparities in Oral Health (often referred to as CAN DO) fluoride varnish trial, a randomized clinical prevention trial (RCT) that evaluated the efficacy of fluoride varnish with oral health counseling to reduce caries incidence in 376 initially caries-free high-risk children aged 6 to 44 months (23, 24). In this article, we used the J-N method for the first time in this fluoride varnish RCT to explore the predictive and prognostic value of the biomarker, salivary mutans streptococci (MS).

This article is organized as follows. In the Methods section, we first describe the standard approach that is used to evaluate treatment-effect heterogeneity. We then show how the J-N method can be used to model baseline characteristics (i.e., covariates) on a continuous scale that retains all of the available information in the covariate. In the results for Applications section, we illustrate how the J-N method can be used to evaluate treatment-effect heterogeneity in the CAN DO fluoride varnish RCT. The Discussion, with further consideration of how to avoid erroneous conclusion and other matters, is presented below. Final remarks are provided in the conclusion.

METHODS

The Standard Approach

The standard approach usually assesses treatment-effect heterogeneity, also called "interactions", "treatment-effect modification" or "moderation", by first defining (often arbitrarily) study-participant subgroups based on categorized (usually dichotomized) baseline characteristics (e.g., low MS levels vs. high MS levels). After the baseline variable is dichotomized, the results are assessed for heterogeneity by testing the statistical interaction between treatment and the dichotomized baseline characteristic for significance. But it is well known that modeling covariates on a continuous scale is more powerful than categorizing covariates into two or more groups (7), since this can diminish the effect of the baseline characteristic as a predictor of treatment effectiveness.

A common mistake is to claim heterogeneity on the basis of separate tests of treatment effects within each level of the baseline variable (2, 3). For example, testing the hypothesis that there is no treatment effect in high MS levels, and then testing it separately in study participants with low MS levels does not address the question of whether treatment differences vary according to MS levels. Another common error is to ignore the uncertainty of the estimates of the treatment effect (e.g., standard error).

The Johnson-Neyman Approach

A more accurate statistical method for assessing the heterogeneity of treatment effects is the Johnson-Neyman (J-N) approach (8–16). Recent developments to this methodology include applications to both longitudinal data with normally distributed continuous outcomes (e.g., hierarchical linear (multi-level models)) (17), as well as non-normally distributed data suitable for generalized linear (mixed) models with adjustment for multiple testing, which protects against erroneous conclusions (18). In Appendix A, we describe how the J-N method can be applied to generalized linear models (GLMs), particularly logistic regression models that assume a logit link and Bernoulli distribution (see Lazar and Zerbe (18) for additional details about the J-N method for generalized linear mixed models). Appendix B provides additional details about the J-N approach that are applicable to the CAN DO fluoride varnish trial described in the Results-Applications section. Appendix C provides a SAS macro based on SAS version 9.3 GLIMMIX procedure that can be used to perform the J-N analysis for logistic regression models with cross-sectional data. It also includes the SAS code used to perform the J-N analysis presented in the Results-Applications section.

To give a brief overview of the method to evaluate heterogeneity of treatment effects, the J-N approach compares the relationship between the treatment effect (e.g., odds ratio) and baseline characteristic (e.g., pre-treatment or pre-intervention variables that do not vary over time) of interest by evaluating whether the fitted regression curves differ significantly among treatment groups, while optionally adjusting for other characteristics. When there is evidence of a heterogeneous treatment effect, then the 'significance region' (the range of the values of the baseline characteristic that significantly differ between the fitted regression curves) can be determined. Determining the significance region involves comparing the treatment curves at each value of the characteristic, which results in statistical tests for every value of the characteristic. In dentistry studies, this can sometimes be well over 1000 tests. Rather than evaluate a potentially infinite number of tests, the J -N analysis can determine the significance region explicitly with adjustment for multiple testing that protects against erroneous conclusions (see Appendix A and Appendix B). While this 'explicit solution' assumes that the patient characteristic (or co-variate) has linear effects, an alternative solution as well as the standard dichotomized approach does not require linearity in the covariates (see Appendix A). A plot of the treatment effect against the baseline characteristic (i.e., covariate) together with the 95% confidence band graphically illustrates heterogeneity of treatment effects.

One advantage of this approach is that by evaluating the prognostic and predictive value of a co-variate, J-N can make more detailed and substantial descriptions than the standard approach. The J-N method makes statements about a significant ‘interaction’ effect and also about the values of the covariate that are significant (e.g., 'There were significant differences between treatments over the whole range of the covariate, and treatment effects increased with increasing value of the covariate'; or possibly, 'Treatment effects increased with increasing value of the covariate, and groups were significantly different for values of the covariate above' (25)). Another advantage over the standard approach that distills heterogeneity of treatment-effects to a not very intuitive p-value is that the J-N method can graphically illustrate the treatment-effects as a function of a continuous covariate.

RESULTS - APPLICATION

Dental caries in children remains a significant public health problem (WHO, 2011). Since caries incidence can be greatest in children and adolescents, dentists need various tools to predict the occurrence of new carious lesions. Some studies have shown that the number of salivary mutans streptococci (MS) is associated with caries onset and progression (26–29), but researchers have questioned whether it is a reliable predictor for dental caries risk in children (30).

Using the J-N method described in the Methods section, we can evaluate the prognostic and predictive value of MS in the CAN DO fluoride varnish study, a controlled dental examiner masked phase III RCT that evaluated the efficacy of fluoride varnish (FV) with oral health counseling versus counseling alone to reduce caries incidence in 376 initially caries-free high-risk children aged 6 to 44 months from predominantly low-income Chinese and Hispanic San Francisco families. The participants underwent dental examination at baseline prior to the intervention and at one and two years post intervention. Intent-to-treat analyses showed a fluoride varnish protective effect in caries incidence, $p<0.001$ (23). Caries incidence was significantly higher for counseling only (0 FV) vs. the groups assigned counseling + FV once/year (OR=2.20, 95% CI: 1.19–4.08) and twice/year (OR=3.77, 95% CI 1.88–7.58))(23).

The study had baseline MS measurements (0 FV per treatment per year: n=89; 1 FV treatment per year n=78; 2 FV treatments per year n=82) for 249 caries-free children. To illustrate the J-N approach more clearly, we collapsed the two FV treatment groups into a single group. We then simultaneously fit two logistic regression curves of any caries incidence, the primary outcome, to evaluate the prognostic and predictive value of MS, which was logarithmically transformed in base 10 ($\log_{10}$ MS). In Figure 1A, the best fitting curves of any caries incidence for the two treatment groups are:

- $-1.15+0.52 \log_{10}$ MS (colony forming units per milliliter or CFU/ml) for no FV, and
- $-1.93+0.33 \log_{10}$ MS (CFU/ml) for FV with associated standard error lines (dashed lines).

Figure 1A summarizes the log odds of caries incidence for FV compared to no FV treatments as MS increases. The figure suggests that higher $\log_{10}$ MS levels were associated with higher log odds of caries incidence, especially for children who were randomly assigned to no FV treatment.

Using the J-N method, we can evaluate the particular values of $\log_{10}$ MS that result in a statistically different (heterogeneous) caries incidence and, most importantly, which children may benefit most from FV treatment. This question may be particularly important, for instance, for parents who are weighing the benefits of their children receiving a treatment against any potential risks. In Appendix B, we show how to use a decision guide to determine whether caries incidence differs between the treatment groups.

Figure 1B displays the relative treatment effectiveness based on odds ratios across individual levels of $\log_{10}$ MS. For high $\log_{10}$ MS values, the estimated odds ratio of caries was lower for children receiving FV than for those not receiving FV (an odds ratio less than one indicates that fluoride varnish was more protective against caries than no FV). The J-N analysis results show evidence for heterogeneity of relative treatment effects (joint test or interaction test: $P<0.0001$, Fig 1B). We conclude that: 1) when $\log_{10}$ MS is between 0.6 and 6.9, children assigned fluoride varnish had significantly lower odds of caries, after adjustment for multiple testing; and 2) children who had the highest values of $\log_{10}$ MS benefitted the most from receiving preventive FV treatment.

To compare the results from the J-N method, we also evaluated treatment-effect heterogeneity using the standard approach, as described in the Methods section. Levels of the biomarker MS were dichotomized as no MS (MS = 0) vs. any (MS >0). We fit a standard logistic regression model of any caries incidence, the primary outcome, to evaluate a treatment by covariate (i.e., dichotomized MS) interaction effect. The interaction test detected borderline statistically significant heterogeneity ($P = 0.06$). We also evaluated MS using the same dichotomization as described in Ramos-Gomez et al.(31) (i.e., $\log_{10}$ MS < 3 vs. log10 MS ≥ 3), and the interaction test between treatment and this dichotomized version of MS detected no statistically significant heterogeneity ($P = 0.21$). SAS version 9.3 GLIMMIX procedure was used to perform both the Johnson-Neyman analysis, as described in Appendix C, and the standard analysis assessing the interaction between treatment and dichotomized baseline characteristic $\log_{10}$ MS.

DISCUSSION

The CAN DO fluoride varnish RCT example illustrates how the Johnson-Neyman analysis addresses research questions that are important for both investigators and parents. Using this method, we were able to evaluate for the first time the heterogeneity of an oral health biomarker, MS, in children enrolled in this study. Our results suggest that children with higher levels of MS who were assigned to receive no fluoride varnish had a poorer prognosis, and that these children may benefit most from fluoride varnish treatment compared with no treatment. This benefit could be explained by the fluoride concentration in saliva that works as the driving force to decrease the rate of enamel demineralization and enhance the rate of mineralization (32–34). Nevertheless, despite the biological plausibility of the observed results, the fluoride varnish protocol did not a priori specify examination of the role of MS as a predictive factor, and these results should be interpreted cautiously. As documented in the primary report of the fluoride varnish trial (23), an unexpected protocol deviation resulted in some children receiving less active fluoride varnish than assigned, and the preventive effects of FV may even be stronger.

The evaluation of treatment-effect heterogeneity has been associated with well-documented problems, such as multiple testing that can lead to erroneous conclusions. However, not presenting evaluations of treatment-effect heterogeneity is a "steep price to pay for a problem that can be remedied by more responsible analysis and reporting." as Lagakos put it (2). To protect against erroneous conclusions that can lead to over-interpreted results, researchers should use a method that accounts for multiple testing, such as Johnson-Neyman. However, the safest approach is to validate exploratory analyses using an independent data set known as "replication" or "external validation". Guidelines have also been proposed to address the challenges associated with the analysis and presentation of the results (3). While these guidelines focus on the reporting of randomized clinical trials, the issues discussed can also apply to observational studies. Therefore, these guidelines are appropriate to apply to any presentation of results that evaluates treatment-effect heterogeneity, including results generated from the Johnson-Neyman approach.

We highlighted in this article how the approach of categorizing baseline characteristics measured on a continuous scale can fail to identify the value of the baseline characteristic as a predictor of treatment effectiveness. While we focused on characteristics measured on a continuous scale, the J-N method can also be applied to other types of numeric characteristics, including discrete count or ordinal (with many categories) scaled baseline characteristics. Subgroup analyses that are properly analyzed and reported can provide useful information for the care of study participants and for future research, but only if the analytic method retains all of the information in the covariate and protects against erroneous conclusions from multiple testing, as the Johnson-Neyman method does. Most importantly, regardless of the analytic method used, exploratory results generated from these subgroup analyses need to be confirmed before the results are applied to patients in the clinical setting. Identification and confirmation of baseline characteristics that predict treatment effectiveness may ultimately lead to personalized oral health care or precision medicine that improves patient health (6).

CONCLUSION

The Johnson-Neyman method enables us to evaluate the prognostic and predictive value of a patient's characteristics. Using methods that predict heterogeneity of treatment effects can improve treatment decisions for individual patients and is becoming an important tool in the field of personalized oral health care.

ACKNOWLEDGEMENTS

This paper benefited greatly from discussion following its presentation to participants at the Methodological Issues in Oral Health Research in Graz Austria organized by Dr. Elisabeth Santigli and from the comments of two anonymous reviewers. We also thank Professor Emeritus Gary Zerbe of the University of Colorado, Denver School of Public Health for his helpful comments. This study was made possible by grant numbers P30-DE-020752, U54 DE014251 and U54 DE019285 from the National Institute for Dental and Craniofacial Research (NIDCR), a component of the National Institutes of Health (NIH). The clinical trial was registered with ClinicalTrials.gov as NCT00066963 and further information can be found at that website.

REFERENCES

1. Assmann SF, Pocock SJ, Enos EE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000; 355:1064–1069. [PubMed: 10744093]
2. Lagakos SW. The challenge of subgroup analyses--Reporting without distorting. New England Journal of Medicine. 2006; 354:1667–1669. [PubMed: 16625007]
3. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine -- Reporting of subgroup analyses in clinical trials. New England Journal of Medicine. 2007; 357:2189–2194. [PubMed: 18032770]
4. Pocock S. More on subgroup analysis in clinical trials. New England Journal of Medicine. 2008; 358:2076–2077. [PubMed: 18463389]
5. Lazar AA, Cole BF, Bonetti M, Gelber RD. Evaluation of treatment-effect heterogeneity using biomarkers measured on a continuous scale: a sub-population treatment effect pattern plot. Statistics in Oncology: Journal of Clinical Oncology. 2010; 28:4539–4544.
6. Garcia I, Kuska R, Somerman MJ. Expanding the foundation for personalized medicine: Implications and challenges for dentistry. Journal of Dental Research. 2013; 92:3S–10S. [PubMed: 23690361]
7. Royston P, Altman D, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: A bad idea. Statistics in Medicine. 2006; 25:127–141. [PubMed: 16217841]
8. Johnson PO, Neyman J. Tests of certain linear hypothesis and their application to some educational problems. Statistical Research Memoirs. 1936; 1:57–93.
9. Johnson PO, Fay LC. The Johnson-Neyman technique, its theory and application. Psychometrika. 1950; 15:349–367. [PubMed: 14797902]
10. Potthoff RF. On the Johnson-Neyman technique and some extensions thereof. Psychometrika. 1964; 29:241–256.
11. Steffens FE. On comparing two simple linear regression lines. South African Statistical Journal. 1968; 2:33–53.
12. Steffens FE. On confidence sets for the ratio of two normal means. South African Statistical Journal. 1971; 5:105–113.

13. Rogosa D. On the relationship between the Johnson -Neyman region of significance and statistical tests of parallel within-group regressions. Educational and Psychological Measurement. 1981; 41:73–84.
14. Hunka S. Identifying regions of significance in AN-COVA problems having non-homogeneous regressions. British Journal of Mathematical and Statistical Psychology. 1995; 48:161–188.
15. Hunka S, Leighton J. Defining Johnson-Neyman regions of significance in the three-covariate AN-COVA using Mathematica. Journal of Educational and Behavioral Statistics. 1997; 22:361–387.
16. Zerbe GO, Archer PG, Banchero N, Lechner AJ. On comparing regression lines with unequal slopes. American Journal of Physiology: Regulatory, Integrative and Comparative Physiology. 1982; 242:178–180.
17. Miyazaki Y, Maier KS. Johnson-Neyman type technique in hierarchical linear models. Journal of Educational and Behavioral Statistics. 2005; 30:233–259.
18. Lazar AA, Zerbe GO. Solutions for determining the significance region using the Johnson-Neyman type procedure in generalized linear (mixed) models. Journal of Educational and Behavioral Statistics. 2011; 36:699–719. [PubMed: 23772174]
19. Huitema, BE. The analysis of covariance and alternatives. New York: Wiley; 1980.
20. Kowalski CJ, Schneiderman ED, Willis SM. Assessing the effect of a treatment when subjects are growing at different rates. International Journal of Biomedical Research. 1994; 37:151–159.
21. Dorsey SG, Soeken KL. Use of the Johnson-Neyman technique as an alternative to analysis of covariance. Nursing Research. 1996; 45:363–366. [PubMed: 8941311]
22. Gillanders BM. Comparison of growth rates between estuarine and coastal reef populations of Achoerodus viridis. Marine Ecology Progress Series. 1997; 146:283–287.
23. Weintraub JA, Ramos-Gomez F, Jue B, et al. Fluoride varnish efficacy in preventing early childhood caries. Journal of Dental Research. 2006; 85:172–176. [PubMed: 16434737]
24. Gansky SA, Cheng NF, Koch GG. Dose-weighted adjusted Mantel-Haenszel tests for numeric scaled strata in randomized trial. Statistics in Biopharmaceutical Research. 2011; 3:266–275. [PubMed: 21709814]
25. Engqvist L. The mistreatment of covariate interaction terms in linear model analyses of behavioral and evolutionary ecology studies. Animal Behavior. 2005; 70:967–971.
26. Loesche WJ, Eklund S, Earnest R, Burt B. Longitudinal investigation of bacteriology of human fissure decay: epidemiological studies in molars shortly after eruption. Infection and Immunity. 1984; 46:765–772. [PubMed: 6500709]
27. Kingman A, Little W, Gomez I, et al. Salivary levels of streptococcus mutans and lactobacilli and dental caries experiences in US adolescent population. Community Dentistry and Oral Epidemiology. 1988; 16:98–103. [PubMed: 3162865]
28. Leverett DH, Proskin HM, Featherstone JD, et al. Caries risk assessment in a longitudinal discrimination study. Journal of Dental Research. 1993; 72:538–543. [PubMed: 8380821]
29. Plonka KA, Pukallus ML, Barnett AG, Walsh LJ, Holcombe TF. A longitudinal study comparing mu-tans streptococci and lactobacilli colonisation in dentate children aged 6 to 24 months. Caries Research. 2012; 46:385–393. [PubMed: 22699390]
30. Thenisch NL, Bachmann LM, Imfeld T, Leisebach MT, Steurer J. Are mutans streptococci detected in preschool children a reliable predictive factor for dental caries risk? A systematic review. Caries Research. 2006; 40:366–374. [PubMed: 16946603]
31. Ramos-Gomez FJ, Weintraub JA, Gansky SA, Hoover CI, Featherstone JD. Bacterial, behavioral and environmental factors associated with early childhood caries. Journal of Clinical Pediatric Dentistry. 2002; 26:165–173. [PubMed: 11878278]
32. Fejerskov O, Thylstrup A, Larsen MJ. Rational use of fluorides in caries prevention. A concept based on possible cariostatistic mechanisms. Acta Odon-tologica Scandinavica. 1981; 39:241–249.
33. Edgar WM, Bowen WH, Amsbaugh S, Monell-Torrens E, Brunelle J. Effects of different eating patterns on dental caries in the rat. Caries Research. 1982; 16:384–389. [PubMed: 6958371]
34. Zero D. In situ caries models. Advances in Dental Research. 1995; 9:214–230. [PubMed: 8615944]

APPENDIX A

The Johnson-Neyman Approach Suitable for Logistic Regression Models

The Johnson-Neyman (J-N) approach was recently developed for the generalized linear model (GLM) and the generalized linear mixed models (GLMM) [1]. The GLM is a special case of GLMM, and it can be extended to the GLMM with the addition of random effects that account for the correlation among longitudinal responses. The methodology for the Johnson-Neyman method suitable for either the GLM and GLMM, which assumes any normal or non-normal distribution from the exponential family, can be found in Lazar and Zerbe [1]. A brief overview is described below about this particular case of the J-N approach for GLM, and this overview is directly applicable to the CAN DO fluoride varnish trial described in the Application Section, particularly logistic regression models of dichotomous outcomes with a logit link and Bernoulli distribution.

Before we provide an overview of the particular case of the J-N approach for logistic regression models for dichotomous outcomes suitable for cross-sectional data, we will first describe how to build a GLM. GLM is a broad class of models suitable for analyzing a diverse type of outcomes (e.g., Bernoulli distribution for dichotomous outcomes or Poisson distribution for counts). (See McCullagh and Nelder [2] and McCulloch and Searle [3] for more information about GLM and GLMM, respectively.) Three decisions are required to build a GLM: the distribution of the data that can be from any probability distribution of the exponential family, the systematic component of the model, and the link function. The distribution can be any member of the exponential family such as Bernoulli, Poisson, negative binomial, or Normal (Gaussian). The systematic component of the model specifies the effect of the covariate, X, on the mean of the distribution, Y. This is known as the linear predictor, $\eta = X\beta$, such that η is linear in β, the regression parameters. The link function, $g(.)$, describes the relationship between the mean of Y or m and η, where $E(Y) = m$ and $g(m) = X\beta$. Any differentiable monotonic link function can be chosen, such as the Bernoulli distribution link function, which is usually the logistic or logit link function, $g(m) = \log(m/(1 - m))$. Logistic regression is one specific kind of GLM with Bernoulli distribution, systematic component(s), X, and logit link function.

Johnson-Neyman Approach for Comparing Logistic Regression Models

Consider two typical logistic regression models fit simultaneously with dichotomous outcomes:

$$\eta_{1j}{=}\log\left(\frac{\Pi_{1j}}{1-\pi_{1j}}\right){=}\alpha_1{+}\beta_1 x_{1j}\eta_{2j}{=}\log\left(\frac{\Pi_{2j}}{1-\pi_{2j}}\right){=}\alpha_2{+}\beta_2 x_{2j}, \quad \text{(A.1)}$$

where x_{ij} denote the measurement of the jth subject for the ith treatment group (here, we consider two treatment groups $i = 1$ or 2 but more than two groups can be considered without loss of generality). Π_{ij} is the probability that the outcome $y_{ij} = 1$, such that y_{ij}~Bernoulli(Π_{ij}), η_{ij} is the linear predictor, and α_i and β_i are the fixed intercepts and slopes for each of the *i* treatment groups. The log $(\frac{\Pi_{ij}}{1-\pi_{ij}})$ expression is usually referred to as the logit or log-odds.

Consider comparing the curve from treatment group 1, ($\alpha_1 + \beta_1 x$), with the curve from treatment group 2, ($\alpha_2 + \beta_2 x$), from models (A.1). The joint hypothesis of equality of the intercepts, α_1 and α_2, and slopes, β_1 and β_2, of two logistic regression curves can be expressed in terms of the difference in the intercepts and the difference in the slopes: $H_k = (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)x_k = \theta_C + \theta_A x_k$, where $\theta_C = (\alpha_1 - \alpha_2)$ and $\theta_A = (\beta_1 - \beta_2)$, for distinct *k* covariate valuesx_1, x_2,x_3,….

In general, when testing a null hypothesis, $\mathrm{H}_0 = \underbrace{\theta}_{sx1} = \underbrace{\begin{pmatrix} \theta_C \\ \theta_A \end{pmatrix}}_{sx1} = \underbrace{C}_{sxp}\underbrace{\beta}_{px1} = 0$, for a fixed contrast matrix, *C*, where β represents the regression parameters, a *t*- or *F*-statistic can be used. When the rank of *C* is one, $t(\hat{\theta}) = \hat{\theta} \Big/ \sqrt{\hat{V}_{\hat{\theta}}}$, where $V_{\hat{\theta}} = CL\hat{C}'$and $\hat{L}$ is the empirical covariance matrix of $\hat{\beta} - \beta$. The null hypothesis can be tested with an *F*-statistic if the rank of *C* is greater than or equal to one, $F(\hat{\theta}) = \left[\underbrace{\hat{\theta}'}_{1xs}\underbrace{(\hat{V}_{\hat{\theta}})}_{sxs}^{-1}\underbrace{\hat{\theta}}_{sx1}\right] \Big/ s$ with *s* numerator degrees of freedom and *v* denominator degrees of freedom. The denominator degrees of freedom are often estimated from the data [4], and in some cases, *v* denominator degrees of freedom may be simply the sample size (n) minus the number of parameters (e.g., for comparing two logistic regression curves as described in A.1, $v = n - 4$). Certainly the usual model assumptions must hold to make valid inference about θ that accounts for the variance-covariance matrix selected by the investigator [5, 6]. A single null hypothesis of the form, $\underbrace{h}_{1xs}\underbrace{\theta}_{sx1} = 0$, can be rejected, $\underbrace{h}_{1xs}\underbrace{\theta}_{sx1} \neq 0$, if $t(h\hat{\theta}) > \sqrt{F_{1-\alpha,1,v}}$, where the *t*-statistic associated with $h\hat{\theta}$ is $t(h\hat{\theta}) = h\hat{\theta} \Big/ \sqrt{h\hat{V}_{\hat{\theta}}h'}$ and $F_{1-\alpha,1,v}$ is the $1 - \alpha$ percentage point of an *F* distribution with 1 and *v* numerator and denominator degrees of freedom, respectively.

Multiple null hypotheses of the form, $h\theta = 0$, can be rejected ($h\theta \neq 0$) whenever the *t*-statistic associated with $h\theta$ exceeds the Scheffé’s criterion, $\sqrt{sF_{1-\alpha,s,v}}$,

$$t(h\hat{\theta}) > \sqrt{sF_{1-\alpha,s,v}} \quad \text{(A.2)}$$

where $F_{1-\alpha,s,v}$ is the $1 - \alpha$ percentage point of an *F* distribution with *s* and *v* numerator and denominator degrees of freedom, respectively. In our application, $\underbrace{h}_{1x2}\underbrace{\theta}_{2x1} = \underbrace{h}_{1x2}\underbrace{C}_{2x4}\underbrace{\beta}_{4x1} = 0$, where

$$\theta=\begin{pmatrix}\theta_C\\ \theta_A\end{pmatrix}=\underbrace{\begin{pmatrix}\alpha_1-\alpha_2\\ \beta_1-\beta_2\end{pmatrix}}_{\theta}=C\beta=\underbrace{\underbrace{\begin{pmatrix}1 & -1 & 0 & 0\\ 0 & 0 & 1 & -1\end{pmatrix}}_{C}\underbrace{\begin{pmatrix}\alpha_1\\ \alpha_2\\ \beta_1\\ \beta_2\end{pmatrix}}_{\beta}}_{\theta=C\beta}$$

, and null hypotheses are therefore rejected whenever $t(h\hat{\theta})>\sqrt{2F_{1-\alpha,2,v}}$. The vertical difference between the two logistic regression equations at any *X* can be expressed as

$$h_x\theta=\underbrace{(1X)}_{h_x}\underbrace{\begin{pmatrix}\alpha_1-\alpha_2\\ \beta_1-\beta_2\end{pmatrix}}_{\theta}=\underbrace{(1X)}_{h_x}\underbrace{\underbrace{\begin{pmatrix}1 & -1 & 0 & 0\\ 0 & 0 & 1 & -1\end{pmatrix}}_{C}\underbrace{\begin{pmatrix}\alpha_1\\ \alpha_2\\ \beta_1\\ \beta_2\end{pmatrix}}_{\beta}}_{\theta=C\beta}=(\alpha_1+\beta_1X)-(\alpha_2+\beta_2X),$$

and multiple null hypotheses can be tested simultaneously at two or more unique values of *X* by repeating A.2 for each value of *X*. Scheffé's method for multiple comparisons controls the family-wise type I error rate for tests of all linear functions, $h_x\theta$, at level α, and the price of multiple comparisons adjustment is always the same no matter how many values of *X* are tested. In our application, Scheffé's constant is $\sqrt{2F_{1-\alpha,2,v}}$, where $s = 2$, which is determined from the number of rows in the *C* matrix, θ is a 2 × 1 vector containing the differences in the intercepts and the differences in the slopes, and $h_x\theta$ are selected to be vertical differences between the lines at specified values of x_k*'s*. This solution that proposes to compare each *X*'s test statistic with Scheffé's criterion does not require linearity in *X*. The above solution is the same solution described in Lazar and Zerbe for GLM and GLMM except in this application we focus on the special case of comparing logistic regression models.

Rather than test a potentially infinite number of hypotheses (one at each x_k) that requires a separate calculation for each x_k, we can determine the significance regions explicitly, as described in Lazar and Zerbe. By re-writing A.2, a quadratic form in *X* is yields $h_x[\theta\theta'-2F_{1-\alpha,2,v}V_\theta]h_x'>0$. If we let $\theta=\begin{pmatrix}\theta_C\\ \theta_A\end{pmatrix}$, where θ_C denote the difference in the intercepts and θ_A denote the difference in the slopes and $V_{\theta(C)}$, $V_{\theta(A)}$ and $\mathrm{Cov}_{\theta(A)\theta(C)}$ denote the empirical error variances (V_θ) and covariance (Cov) of θ_C and θ_A, then the quadratic inequality is:

$$(1X)\left[\begin{pmatrix}\theta_C^2 & \theta_C\theta_A\\ \theta_A\theta_C & \theta_A^2\end{pmatrix}-2F_{1-\alpha,2,v}\begin{pmatrix}V_{\theta(C)} & \mathrm{cov}_{\theta(C)\theta(A)}\\ \mathrm{Cov}_{\theta(A)\theta(C)} & V_{\theta(A)}\end{pmatrix}\right]\begin{pmatrix}1\\ X\end{pmatrix}>0 \text{ or}$$

$Ax^2+Bx+C > 0$, where $A=\theta_A^2-2F_{1-\alpha,2,v}V_{\theta(A)}$, $B = 2(\theta_A\theta_C - 2F_{1-\alpha,2,V}\mathrm{Cov}_{\theta(A)\theta(c)})$, $C=\theta_C^2-2F_{1-\alpha,2,v}V_{\theta(C)}$, $D = B^2 - 4AC$. Here we focus on two treatments groups, but more than two groups can be compared.

After estimating A, B, C and D, one of three possible cases is used to determine the significance regions explicitly.

Case I

If A>0 (implies D>0), then the tests are significant for the *X*'s satisfying: $X<\left[\frac{-\mathrm{B}-\sqrt{D}}{2\mathrm{A}}\right]$ or $X>\left[\frac{-\mathrm{B}-\sqrt{D}}{2\mathrm{A}}\right]$.

Case II

If A<0 and D>0, then the tests are significant for the *X*'s satisfying:

$\left[\frac{-\mathrm{B}+\sqrt{D}}{2\mathrm{A}}<X<\frac{-\mathrm{B}-\sqrt{D}}{2\mathrm{A}}\right]$.

Case III

If D<0 and A<0, then the tests are never statistically significant for any *X*. The case where A>0 and D<0 cannot occur.

The 'explicit' solution described above is the same solution described in Lazar and Zerbe [1], which assumes that the two regression functions are linear in *X*. Additional clarification about the cases is described in Lazar and Zerbe.

APPENDIX B

J-N decision guide for determining the significance regions with application to the CAN DO fluoride varnish trial

We illustrate how a decision guide useful for determining the appropriate case can be used in the fluoride varnish study (Figure 2). This guide provides a facile way to evaluate whether the curves are statistically different with adjustment for multiple comparisons and without calculation of components A, B, C, or D. In addition, this decision guide supplies additional tests to validate the selection of the appropriate case.

The first step of the decision guide evaluates the joint test, which is a simultaneous test of the equality of the intercepts and the equality of the slopes. If the joint test is not statistically significant, then the lines do not differ in either intercepts or slopes. This is Case III, and it is not necessary to calculate A, B, C, or D. If the joint test is statistically significant, as in the fluoride varnish (FV) study (*F*-value = 10.15, s = 2, v = 243; P<0.0001), then D is greater than zero, which agrees with the results from the FV study. The next step tests whether the equality of the slopes using Scheffé's criterion ($2F_{1-\alpha,2,\nu}$) is statistically significant. Case I (A>0) occurs if, and only if, the slopes differ by Scheffé's criterion. Case II (A<0 and D>0) occurs when the slopes do not differ by Scheffé's criterion, and the simultaneous test of equality of the intercepts and slopes is significant by the F-test.

For the fluoride varnish study, the estimates of A, B, C and D are a = −0.1023, b = +0.81356, c = −0.48512 and d = +0.46389, respectively. Since a<0 (implies d >0), we then select Case II. Case II is verified by checking whether the slopes do not significantly differ (p > 0.05) using Scheffé's criterion, where the |t-value| < Scheffé's criterion:

$\left(1.29<2.47=\sqrt{(2*3.03)}\right)$. After plugging in a, b, c and d for Case II, as described in Appendix A, the results are $0.649 < \log_{10}$MS (CFU/ml)<7.324. We restrict our inference to the MS range where the data were collected. Therefore, we conclude that the results are $0.649 < \log_{10}$ MS (CFU/ml)< 6.99.

As shown for the fluoride varnish study, the explicit solution eliminates testing at every value of the covariate. For two or more covariates of interest, this explicit solution is thought to be intractable [7,8]. Hunka and Leighton [9] developed a solution for three covariates by casting an equation within a general linear model framework, which cannot be solved directly. Symbolic processing capabilities of computational software, such as Mathematica, can be used to provide a solution. Hunka and Leighton's solution is limited to the analysis of covariance framework for non-correlated data and it does not consider the problem of evaluating treatment-effect heterogeneity for baseline characteristics, as described in this paper.

APPENDIX C

SAS macro for evaluating treatment-effect heterogeneity using the Johnson-Neyman methodology

The following SAS macro was used to analyze the CAN DO fluoride varnish data using the Johnson-Neyman logistic regression analysis of any caries incidence vs. no caries incidence (outcome), a dichotomous outcome with a continuous covariate of interest, MS.

The SAS Code starts on the next page.

```
/***************************START SAS CODE*****************************/
/* SignificanceRegion.sas
Macro for performing Johnson-Neyman for a logistic regression model.
(Similar SAS code adapted from the generalized linear mixed model de-
scribed by Lazar and Zerbe 2011)
The input parameters are:
SignificanceRegion
datasetname = data set name
outcome = Dichotomous outcome variable (0 or 1)
treatment = Treatment Indicator (0 or 1)
continuousvar= baseline continuous variable of interest
*/
%macro SignificanceRegion(datasetname,outcome,treatment,continuousvar);
proc glimmix data=&datasetname oddsratio ;
class &treatment;
model &outcome(event='1')= &treatment &treatment*&continuousvar /s
dist=bin link=logit noint;
/*Joint Test*/
```

```
contrast 'contrast Fixed' &treatment 1, &treatment*&continuousvar
1 ;
/*Difference in Slopes to Calculate A*/
estimate 'slope' &treatment*&continuousvar −1 1;
/*Difference in the Intercepts to Calculate C*/
estimate 'int' &treatment −1 1 ;
/*Used to Calculate the Covariance for B*/
estimate 'intslope' &treatment −1 1 &treatment*&continuousvar −1 1;
/* Outputs the above information using the output delivery system*/
ods output estimates=estimates;
/*Alternative way to determine the Significance Region is to compare
Scheffe's Constant to the t-value based on the following calculation
/*
estimate 'check 0' &treatment 1 &treatment *&continuousvar 0/or;
estimate 'check 3.5' &treatment 1 &treatment *&continuousvar 3.5/or;
estimate 'check 6' &treatment 1 &treatment *&continuousvar 6/or;
*/
run;
proc sort data=estimates;by df;
data f;set estimates;by df;
/*Denominator degrees of freedom determined from the joint test speci-
fied in the model*/
if first.df;
/*F-statistic*/
F=finv(0.95,2,df);
/*Use Scheffe to compare to the t-statistic*/
Scheffe = sqrt(2*f);
match=1;
keep f match Scheffe;
run;
proc print data=f;run;
data estimates2;set estimates;match=1;run;
data estimates3;merge f estimates2 ;by match;
data intestimates;set estimates3;
if label='int' then do;InterceptEst=estimate;IntVar=StdErr**2;
C = (InterceptEst**2)−(2*f*IntVar);output;end;
keep C InterceptEst IntVar f;
run;
data slopeestimates;set estimates3;
if label='slope' then do;SlopeEst = estimate;SlopeVar=StdErr**2;
slopesq=SlopeEst**2;
A = (slopesq)−(2*f*SlopeVar);output;end;
keep A SlopeEst SlopeVar ;
run;
```

```
data intslope;set estimates3;if label='intslope' then do;IntSlopeEst =
estimate;IntSlopeVar=StdErr**2;output;end;
keep f IntSlopeEst IntSlopeVar;
run;
data all;merge intestimates slopeestimates intslope;
/* could calcuate cov using formula as described in Lazar & Zerbe
(2011);
Form 1: V((a1−a2)+(b1−b2))=V(a1−a2)+V(b1−b2)+2Cov((a1−a2),(b1−b2))
So Cov((a1−a2),(b1−b2)) =( V((a1−a2)+(b1−b2))−V(a1−a2)−V(b1−b2))/2
Or use this formula; same formula as described in Lazar & Zerbe (2011);
Form 2: V((a1−a2)−(b1−b2))=V(a1−a2)+V(b1−b2)−2Cov((a1−a2),(b1−b2))
So Cov((a1−a2),(b1−b2))=(V(a1−a2)+V(b1−b2)−V((a1−a2)−(b1−b2)))/2*/
covariance=(IntSlopeVar-intvar-slopevar)/2; /*using form 1*/
run;
data all1;set all;
B=2*((InterceptEst*SlopeEst)−(2*f*Covariance));
D=(B**2)−(4*A*C);
f2sqrt=sqrt(f*2);
if A>0 then do; Case1lower=((−B−sqrt(D))/(2*A));Case1upper = ((−B+sqrt
(D))/(2*A));end;
if A<0 and D>0 then do; Case2lower=((−B+sqrt(D))/(2*A)); Case2upper=((−B
−sqrt(D))/(2*A));end;
label Case1Lower= 'Case I X< ';
label Case1Upper ='Or X>';
label Case2Lower = 'Case II';
label Case2Upper = '<X< ';
F_To_Compare_to_T=sqrt(f*2);
label F_To_Compare_to_T = 'Scheffe Constant For Comparing to T-Value';;
run;
proc print data=all1 noobs label;var A B C D Case1Lower Case1Upper
Case2Lower Case2Upper F_To_Compare_to_T;run;
%mend;
/*
Macro call for CANDO fluoride varnish dataset
Intended is the name of the dataset, dfs1 is the outcome variable,
intended1 is the treatment variable, mutansStrep is the variable of in-
terest measured on a continuous scale;
*/
%SignificanceRegion(intended,dfs1,intended1,mutansStrep);
/*************************END SAS CODE*********************************/
```

Before using the above macro, we suggest that the appropriate model be fitted to the data using the usual model fitting criteria. Examples of the Johnson-Neyman approach for longitudinal data suitable for generalized linear mixed models using SAS GLIMMIX with associated SAS code can be found in Lazar and Zerbe (1).

REFERENCES

1. Lazar AA, Zerbe GO. Solutions for determining the significance region using the Johnson-Neyman type procedure in generalized linear (mixed) models. Journal of Educational and Behavioral Statistics. 2011; 36:699–719. [PubMed: 23772174]
2. McCullagh, P.; Nelder, JN. Generalized linear models. 2nd ed.. London: Chapman and Hall/CRC Press; 1989.
3. McCulloch, C.; Searle, S. Generalized, linear and mixed models. New York: John Wiley and Sons, Inc; 2001.
4. Molenberghs, G.; Verbeke, G. Models for discrete longitudinal data. New York: Springer; 2005.
5. Littell, RC.; Milliken, GA.; Stroup, WW.; Wolfinger, RD.; Schabenberger, O. SAS for mixed models. 2nd ed.. Cary: SAS Institute Inc; 2006.
6. Snedecor, GW.; Cochran, WG. Statistical methods. 8th ed.. Ames: Iowa State University Press; 1989.
7. Johnson PO, Fay LC. The Johnson-Neyman technique, its theory and application. Psychometrika. 1950; 15:349–367. [PubMed: 14797902]
8. Miyazaki Y, Maier KS. Johnson-Neyman type technique in hierarchical linear models. Journal of Educational and Behavioral Statistics. 2005; 30:233–259.
9. Hunka S, Leighton J. Defining Johnson-Neyman regions of significance in the three-covariate ANCOVA using Mathematica. Journal of Educational and Behavioral Statistics. 1997; 22:361–387.

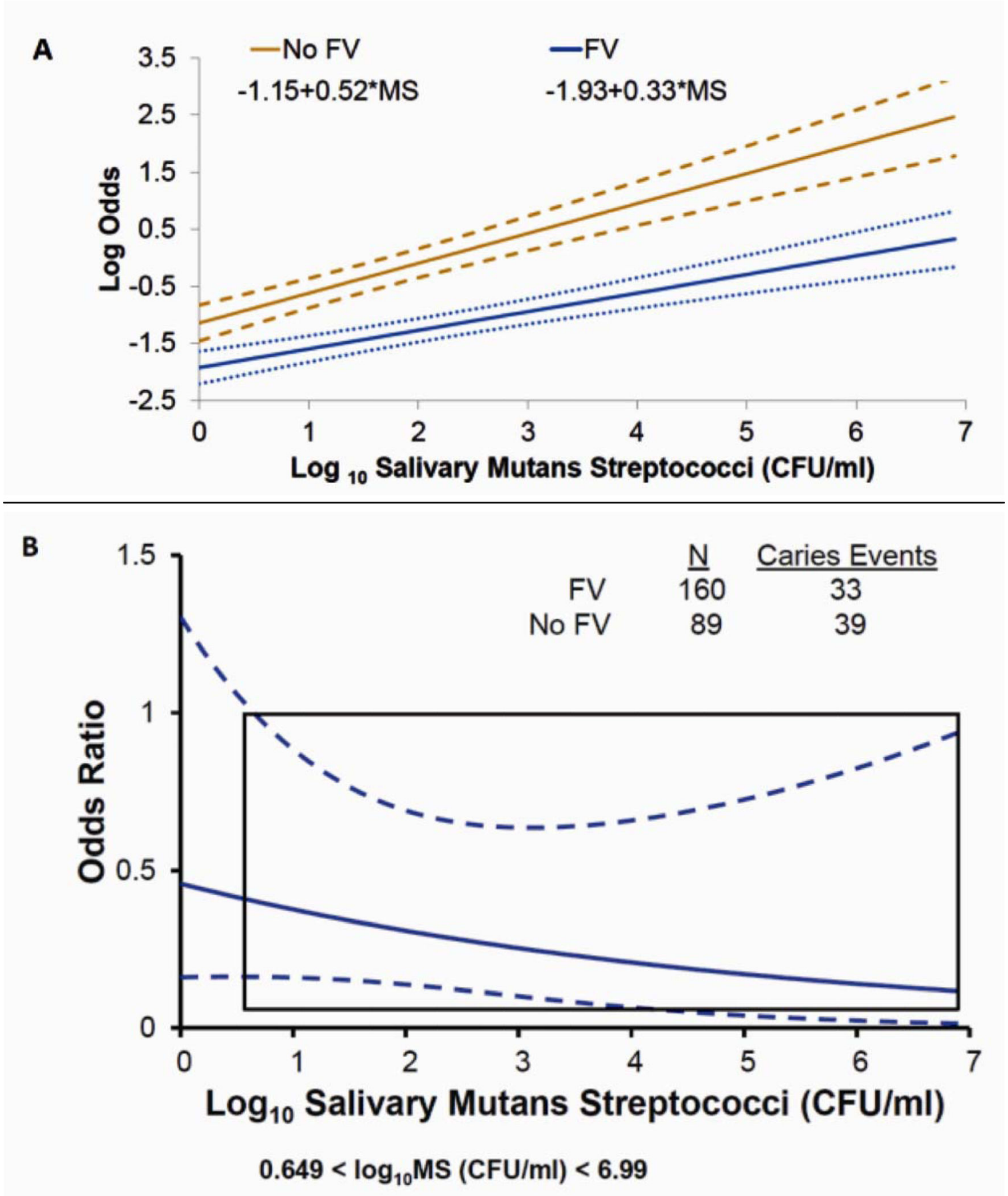

Figure 1.

Johnson-Neyman analysis of the treatment effect of fluoride varnish (FV) versus no fluoride varnish (No FV) treatment as measured by **(A)** log odds of caries incidence with standard error lines (dashed lines) **(B)** odds ratio (less than one indicates FV better; otherwise No FV better) with corresponding 95% confidence band using Scheffé's method (dashed lines). The x-axis indicates the values of the children's Log_{10} salivary mutans streptococci (CFU/ml).

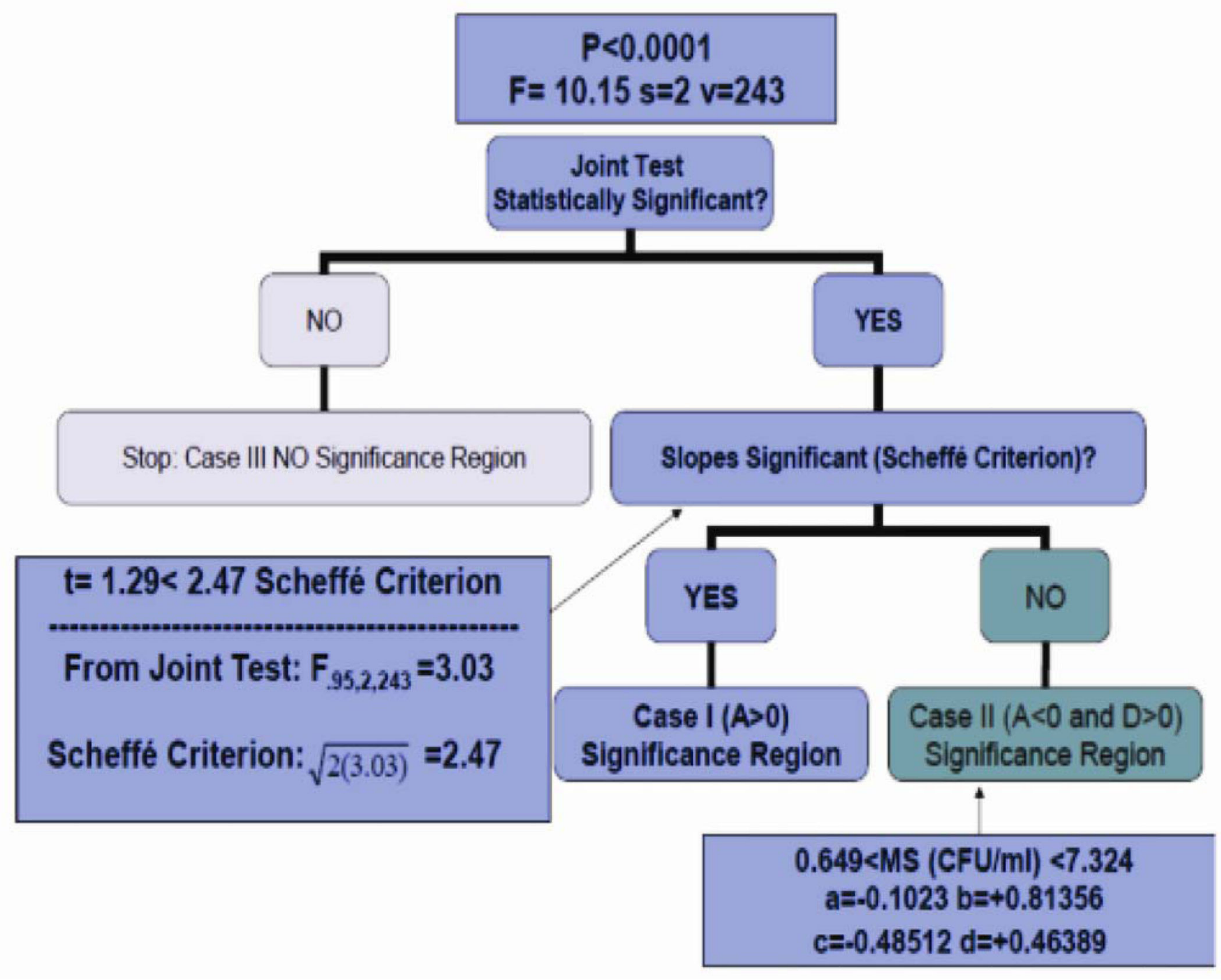
P<0.0001
F= 10.15 s=2 v=243
Joint Test Statistically Significant?
NO
YES
Stop: Case III NO Significance Region
Slopes Significant (Scheffé Criterion)?
t= 1.29< 2.47 Scheffé Criterion
From Joint Test: F.95,2,243 =3.03
Scheffé Criterion: √2(3.03) =2.47
YES
NO
Case I (A>0) Significance Region
Case II (A<0 and D>0) Significance Region
0.649<MS (CFU/ml) <7.324
a=-0.1023 b=+0.81356
c=-0.48512 d=+0.46389

Figure 2.

Decision guide applied to the fluoride varnish study described in Application section.