Theses and Dissertations--Epidemiology and Biostatistics

College of Public Health

2017

An Exploratory Statistical Method For Finding Interactions In A Large Dataset With An Application Toward Periodontal Diseases

Joshua Lambert
University of Kentucky, joshua.lambert@uky.edu

Digital Object Identifier: https://doi.org/10.13023/ETD.2017.448

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student’s advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student’s thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Joshua Lambert, Student

Dr. Heather Bush, Major Professor

Dr. Steve Browning, Director of Graduate Studies

An Exploratory Statistical Method For Finding Interactions In A Large Dataset With An Application Toward Periodontal Diseases

DISSERTATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Public Health at the University of Kentucky

By
Joshua Lambert
Lexington, Kentucky

Director: Dr. Heather Bush, Associate Professor of Biostatistics
Lexington, Kentucky 2017

ABSTRACT FOR DISSERTATION

An Exploratory Statistical Method For Finding Interactions In A Large Dataset With An Application Toward Periodontal Diseases

It is estimated that Periodontal Diseases effects up to 90% of the adult population. Given the complexity of the host environment, many factors contribute to expression of the disease. Age, Gender, Socioeconomic Status, Smoking Status, and Race/Ethnicity are all known risk factors, as well as a handful of known comorbidities. Certain vitamins and minerals have been shown to be protective for the disease, while some toxins and chemicals have been associated with an increased prevalence. The role of toxins, chemicals, vitamins, and minerals in relation to disease is believed to be complex and potentially modified by known risk factors. A large comprehensive dataset from 1999-2003 from the National Health and Nutrition Examination Survey (NHANES) contains full and partial mouth examinations on subjects for measurement of periodontal diseases as well as patient demographic information and approximately 150 environmental variables. In this dissertation, a Feasible Solution Algorithm (FSA) will be used to investigate statistical interactions of these various chemical and environmental variables related to periodontal disease. This sequential algorithm can be used on traditional statistical modeling methods to explore two and three way interactions related to the outcome of interest. FSA can also be used to identify unique subgroups of patients where periodontitis is most (or least) prevalent. In this dissertation, FSA is used to explore the NHANES data and suggest interesting relationships between the toxins, chemicals, vitamins, minerals and known risk factors that have not been previously identified.

KEYWORDS: Biostatistics, Statistics, Feasible Solution, Algorithm, Big Data, Interaction

Author's signature: Joshua Lambert

Date: November 29, 2017

An Exploratory Statistical Method For Finding Interactions In A Large Dataset With An Application Toward Periodontal Diseases

By

Joshua Lambert

Director of Dissertation: Heather Bush

Director of Graduate Studies: Steve Browning

Date: November 29, 2017

To Jacob, Sarah, and Rebekah

ACKNOWLEDGMENTS

As I finish up my 11th year of college, there are many people who I would like to acknowledge.

Thank you to my friends. I can honestly say that I have the greatest group of friends a person could ask for. Your support, guidance, and mentorship has made me into who I am today. Thank you for taking the time to explain the simple things, have lunch, go on trips, laugh, cry, and deal with all my antics.

Thank you to all my teachers, pastors, and mentors. To my fourth grade remedial math teachers for not giving up on me and taking the time to show me that math is fun. To Dr. Chris Mecklin for introducing me to college level statistics. Your classes will always be my favorite undergraduate classes. Thank you to Mark Randall and Ryan Brooks at Murray State Chi-Alpha. You brought me in my Freshman year of college and treated me like a son and brother. I love you guys. Thank you to all my committee members for all of your support and feedback. To Dr. Heather Bush for her friendship, mentorship, and introducing me to my dissertation topic and being my advisor. Thank you for teaching me what it meant to be a collaborative statistician and a good consultant. To Dr. Arny Stromberg for his friendship, mentorship, and introducing me to FSA. Thank you for listening to me complain! Also, Thank you Arny for giving me a job that enabled me to support my family while I finished.

Thank you to all my family for always asking me how school was going and being so supportive. To my Aunt Ginny and my Dad for always letting me know how proud they are. I'm so proud to be a Lambert! To my Mom and Step Dad. Your love for me is something I never doubted. Thank you for everything you have given up for me. You showed me what it meant to work hard, be patient, and do what was best for family and friends. Thank you for your willingness to take on such a big role with

Sarah and Jacob over the last year so that this dissertation could be completed. I love you guys.

To my kiddos, Jacob and Sarah. You both have been my sunshine these last 3 years. You have been the source of so much joy in both your Mothers' and my life. I love you so much.

To my Wife, Rebekah. As I write this I can't help but think about how hard this has been for the both of us. It seems like this has taken an eternity. Your smile and support has been what has kept me going. Thank you for always being so excited for me and supportive of me chasing my dreams. You have been there every step of the way and made sure that I always was taken care of and loved. Thank you, I love you so much.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction to Periodontal Diseases

The leading cause of tooth loss in adults is periodontal disease [1][2] [3]. Both tooth loss and periodontitis are associated with a decline in social, psychological, and environmental quality of life measurements in adults and the elderly [4] [5] [6]. Bad breath, red or swollen gums, tender or bleeding gums, painful chewing, loose teeth, and sensitive teeth are all clinical signs of the disease [7]. Known risk factors include smoking, diabetes, poor oral hygiene, stress, heredity, and malocclusion [7]. Because poor oral health is being linked to systemic disease status, monitoring the severity of periodontitis in the adult U.S. population is now a part of the health-promotion and disease-prevention activities of Healthy People 2020 [8].

Periodontal diseases are a result of an infection or inflammation of the gums caused by pathogenic microorganisms in the dental biofilms (plaque) that form on the teeth and in the subjingival sulcus [9] [10]. This infection causes gingival soft tissue damage resulting in pockets or deepened crevices between the gingiva and tooth root [10]. In its earliest stage, gingivitis, the gums can become swollen and red, and may bleed. This can progress to the gums pulling away from the tooth, enhancing the opportunity for alveolar bone loss and leading to a loose tooth or even exfoliation [9]. The progression of the inflammatory lesion through this process with bone loss is called periodontitis and ranges in severity from mild to severe (definitions below) [11].

1.1 Periodontal Prevalence

Due to the inconsistencies in measurement of periodontitis (see periodontal surveillance section), it is difficult to accurately describe the prevalence of the disease worldwide [12]. Countries have utilized different measurements and case definitions, and measurements used to diagnose periodontitis have changed over time, which has made it difficult to compare prevalence. The diagnosis of periodontal disease is now based on the measurements of periodontal pockets via calibrated probing(probing pocket depth or PPD) , loss of clinical attachment (clinical attachment level or CAL), the pattern and extent of bone loss, or a combination of these measures. PPD is the distance from the gingival margin to the base of the gingival sulcus or periodontal pocket. CAL is the distance from the cemento-enamel junction (CEJ); or another definite chosen landmark to the base of the sulcus or periodontal pocket. Since case definitions for periodontitis have changed over time, its extremely important that the measurement of PPD and CAL are accurate and reproducible.

1.2 Periodontal Disease Classification

Case definitions for periodontal disease types were re-examined and classified by the Center for Disease Control (CDC) and American Academy of Periodontology (AAP) in 1999. The next update for classifying periodontal diseases is scheduled for 2017. There are eight separate classifications of periodontal disease: gingival disease, chronic periodontitis, aggressive periodontitis, periodontitis as a manifestation of systematic disease, necrotizing periodontal disease, abscesses of the periodontium (the tissues that invest and support the teeth including the gingiva, alveolar mucosa, cementum, periodontal ligament, and alveolar supporting bone), periodontitis associated with endodontic lesions, developmental or acquired deformities and conditions.

1.3 Definitions for Periodontal Diseases

Gingival disease, or gingivitis, “is the mildest form of periodontal disease. It causes the gums to become red, swollen, and bleed easily. There is usually little or no discomfort at this stage. Gingivitis is often caused by inadequate oral hygiene. [13]” . Chronic periodontitis “results in inflammation within the supporting tissues of the teeth, progressive attachment and bone loss. This is the most frequently occurring form of periodontitis and is characterized by pocket formation and/or recession of the gingiva. It is prevalent in adults, but can occur at any age. Progression of attachment loss usually occurs slowly, but periods of rapid progression can occur." [13]. Aggressive periodontitis "occurs in patients who are otherwise clinically healthy. Common features include rapid attachment loss and bone destruction and familial aggregation." [13] Periodontitis as a manifestation of systematic " conditions such as heart disease, respiratory disease, and diabetes are associated with this form of periodontitis." [13] Necrotizing periodontitis "is an infection characterized by necrosis of gingival tissues, periodontal ligament and alveolar bone. These lesions are most commonly observed in individuals with systemic conditions such as HIV infection, malnutrition and immunosuppression." [13] Abscesses of the periodontium is defined as "a localized collection of pus (i.e. an abscess) within the tissues of the periodontium. It is localized purulent collection in the periodontal tissues." [14] Periodontitis associated with endodontic lesions is a "bacterial infection from a periodontal pocket associated with loss of attachment and root exposure may spread through accessory canals to the pulp, resulting in pulpal necrosis." [13]. Lastly, periodontal disease associated with developmental or acquired deformities and conditions can be one of the following: localized tooth-related factors that modify or predispose to plaque-induced gingival diseases/periodontitis, mucogingival deformities and conditions around teeth, or occlusal trauma.

1.3.1 Periodontitis Prevalence Classified by Severity

Both chronic and aggressive periodontitis are further classified by severity. First, moderate and severe periodontitis are defined based on measurements of PPD and CAL at interproximal sites. Severe periodontitis requires two or more interproximal sites with CAL 6 mm, not on the same tooth, and one or more interproximal sites with PPD 5 mm. Moderate periodontitis is defined as two or more interproximal sites with CAL 4 mm, not on the same tooth, or two or more interproximal sites with PPD 5 mm, not on the same tooth. However, this definition was not inclusive of any mild definition and therefore gave incomplete estimations for the prevalence of periodontitis in the United States. In 2012, the CDC and AAP added a case definition for mild periodontitis. Mild Periodontitis is defined as 2 interproximal sites with AL 3 mm, and 2 interproximal sites with PPD 4 mm (not on same tooth) or one site with PPD

5 mm [15]. Total periodontitis is now defined as the presence of severe or non-severe periodontitis (mild and moderate) [15]. The mild stage of periodontitis is estimated to affect between 50-90% of adults worldwide [16]. Advanced stages of periodontitis occur less frequently, estimated to be less than 10-15% [17]. Periodontitis is less common in children across all populations [18]. The extent of periodontitis within the US adult population was examined in the 2009-2012 cycles of the NHANES survey. During that time period, 46% of US adults had some form of periodontitis while, 9% were classified as having severe periodontitis [15]. In this chapter, further information about periodontal etiology, case definitions, risk factors, and its relationship to other diseases will be outlined. Also, current surveillance efforts will be detailed.

1.4 Etiology of Periodontitis

Chronic periodontitis is the most prevalent form of periodontitis encountered dental in practice [19]. In terms of the etiology of the disease, periodontal bacterial biofilms accumulate related to diet, behaviors, and oral hygiene practices that drive periodontal disease initiation and progression [20]. Although, individuals with plaque build-up will not necessarily develop periodontitis, the bacterial accumulation plays a required role in the development of periodontitis which trigger biological mechanisms contributing to the development of periodontitis. For example, microbiological studies of the disease have identified the ”Red-Complex” of bacteria implicated for the initiation and progression of chronic periodontitis. This complex is composed of specific bacteria species, including *Porphyromonas gingivalis*, *Treponema denticola*, and *Tannerella forsythia* [20] [21]. Importantly, recent studies have shown that co-operative or synergistic effects among these species, can elicit greater tissue and bone loss than the individual infections [22].

1.4.1 Host Immunity

Chronic inflammation of the periodontal tissue is a result of bacterial colinization of mucosal cells and invasion into deeper preiodontal tissue, and is considered the proximate cause of the tissue destruction in periodontitis. Since chronic inflammation induced by pathogenic biofilms causes periodontitis, it would be expected that periodontal pathogens activate or suppress aspects of the immune system leading to a dysregulated immune system and chronic inflammation[23]. The current paradigm explaining this phenomenon indicates that the periodontal pathogens have an enhanced capacity to invade cells, altering their functions, as well as biology of many members of the oral microbiome and is the key event in the initiation of periodontitis [24]. The persistence of this bacteria within these tissues altering the local microenvironment and physiology of the oral microbiome drives the chronic inflammation. Thus, while the quantity and quality of bacterial accumulation juxtaposed to gingival tissues is a key component to the initiation and progression of periodontitis [24]; the host inflammatory and immune response cells and factors that are activated directly result in collateral damage. [25] [26] [27] [28].

1.5 Risk Factors of Periodontitis

1.5.1 Demographics

Age: Periodontal disease prevalence increases with age, and ages effect has been defined as a cumulative in nature [29]. The estimated prevalence of total periodontal disease for US residents from the 2009-2012 National Health and Nutrition Examination (NHANES) full mouth examination was 24.8%, 37.2%, 52.7%, and 68.0% for those aged 30-34, 35-49, 50 to 64, and 65 and older, respectively [30]. The overall prevalence of periodontitis for U.S dentate adults age 30 or older was 46%. This data therefore indicates increased prevalence of periodontitis with age.

Sex: Men consistently show a higher prevalence and severity of periodontal diseases than women. Although, after onset, periodontal progression appears to be similar in both men and women [31]. The NHANES 2009-2012 data showed the same differences between men and women. 54.9% of men had periodontitis, while only 37.4% of women had some form of periodontitis [30]. Men commonly have a heightened innate immune response system which also may contribute to the increased risk of destructive periodontal disease [31].

Race/Ethnicity: Racial/ethnic minorities are at the greatest risk for periodontal disease. For adults age 30 and over it is estimated that, 63.5% of Hispanic-Americans, 59.1% of Non-Hispanic Blacks, and 50% of Non-Hispanic Asian Americans are affected by periodontitis based on NHANES data from 2009 to 2012. [15]. For adults age 30 and over, 42.6% of Non-Hispanic Whites are estimated to be affected by periodontitis based on the 2009 to 2010 NHANES data [32].

Socioeconomic and Educational Status: Income and educational status have also been shown to be associated with periodontal disease prevalence [33] [34] [15]. One cross-sectional study at Narayana Dental College and Hospital examined the education level and its correlation to periodontal status. Those without a high school education were estimated to be as much as twice as likely to have periodontal disease [34]. In NHANES data 2009-2012, it was estimated that those with low socioeconomic status had twice the prevalence of periodontal disease compared to those with high socioeconomic status [15].

1.5.2 Genetic Factors

In recent years, genetic variations and mutations that can alter the host immune response and may link with periodontal disease expression have also been identified. Genes encoding TNF, IL-1, FCRIIIb NA1/NA2, HLA-A9, and others have all been associated with aggressive periodontitis [35] [36] [37] [38]. Specific polymorphisms in the IL-1β gene and resulting response levels have been associated with chronic periodontitis[38]. These findings were extended such that utilizing this reported genetic predisposition, a periodontal susceptibility test (PST) was developed to identify individuals at increased risk for chronic periodontitis [39]. The marketing of this PST test and ongoing research to identify additional genetic contribution to disease focused on providing clinicians the ability to more accurately identify individuals susceptible to periodontal disease and initiate improved preventative measures.

1.5.3 Modifiable Risk Factors

A modifiable risk factor is a risk factor, such as lifestyle, that an individual can take measures to change and subsequently impacts their risk profile. Smoking tobacco is the strongest modifiable risk factor identified for periodontitis [40] [41]. The observed effects are dose-dependent; as the number of years of smoking increases, so does the severity of periodontitis [42]. Tobacco smoke toxins increase the levels of pro-inflammatory cytokines, as well as dysregulates neutrophil functions in response to microbial challenge [43]. Not only has smoking been implicated as a major risk factor for periodontitis, it has also been shown to impede healing after periodontal therapy [44]. Another example of a modifiable risk factor for periodontitis is alcohol consumption with a dose-dependent relationship with periodontitis and alveolar bone loss [45].

The understanding of how external or environmental factors effect molecular or genetic expression within patients may provide further insights into the understanding of periodontal disease [46]. Other inflammatory diseases have been shown to be epigenetic in nature [46].

1.5.4 Environmental factors

Environmental factors can be considered modifiable as the exposure to toxins in an environment or nutrients in a diet could be altered. Various chemical toxins, allergens, bacteria, and nutrients have been associated with periodontitis in the past [47] [48], albeit research examining chemical environmental toxins and their role in the development in periodontal disease is limited. However, as examples, cadmium and lead have both been associated with an increased likelihood for periodontitis in men and women [48] [49]. In terms of dietary nutrients, increased intake of foods high in Vitamins A, E, and C have generally been reported to decreased the severity of periodontitis [50]. Dietary minerals (calcium, iron, etc) can also modulate the effects of periodontal disease extent or have been related to improved response to treatment [51]. More details on these environmental associations will be explored in Chapter 2 and Chapter 4.

1.6 Systematic Disease and Periodontal Disease

In 2016, a review for clinicians was released by the Indiana Dental Association to present the most recent literature discussing systematic conditions that identify periodontal disease as a risk factor. These conditions include: cardiovascular disease, cerebrovascular diseases, peripheral arterial disease, respiratory diseases, low birth weight, Type 2 diabetes mellitus, insulin resistance, rheumatoid arthritis, obesity, osteoporosis, and complications of pregnancy [52]. This report, as well as many others, have observed most frequently a link with cardiovascular disease and diabetes mellitus [53] [54]. The following subsection will focus on these two systemic diseases and how they relate to periodontal disease.

1.6.1 Cardiovascular Disease

The main form of cardiovascular disease, atherosclerosis, has been strongly correlated with periodontal disease. Atherosclerosis occurs when fatty deposits, and the lipid and inflamed plaques clog the arteries. These plaques are composed of cholesterol, fatty substances, cellular waste products, calcium and fibrin (a clotting material in the blood) [55]. The walls of the artery harden, become narrower, and subsequently restrict blood flow as the plaque accumulates on the walls of the arteries and veins. Evidence has linked periodontal bacterial infections to athlerosclerosis [56] [57] [58]. Patients with periodontal disease have a higher incidence of cardiovascular disease and atherosclerotic and cardiovascular events [59] [60] [61] [62]. It is theorized that chronic periodontal inflammation, exacerbates a systemic inflammatory challenge, leading to an increase in cytokines, which may directly contribute to athlerosclerosis [56] [57] [58]. Periodontitis also leads to increased levels of C-reactive protein, IL-6, and neutrophil inflammatory products[63] that may contribute to the biology of athlerosclerosis [64]. Furthermore, periodontal bacteria (pathogens) have been reported to be localized in atherosclerotic lesions in diseased individuals [64]. While the link between atherosclerosis and periodontal disease is compelling, evidence demonstrating a causal link still needs to be delineated. The association is strong and evident, but this simply could be exemplifying shared risk factors that predispose an individual to both periodontal disease and athlerosclerosis (age, diet, smoking, ect) [65].

1.6.2 Diabetes Mellititus

The relationship between diabetes mellitus and periodontitis is considered bi-directional [54], with diabetes as a major risk factor for periodontitis, particuliarly Type 2 diabetes (T2DM). [66] [67] [68]. Individuals are estimated to have three times the risk for periodontitis if they have diabetes compared to not having diabetes [69]. Tsai et al. found that NHANES III subjects with poorly controlled diabetes had a significantly higher prevalence of severe periodontitis than those without diabetes [70]. Studies of Type I diabetes (eg. T1DM, juvenile diabetes) has found that 10% of children with T1DM displayed periodontal symptoms (attachment loss and bone loss) despite having similar levels of bacterial accumulation and oral plaque indices as controls [71]. While the risk of developing periodontitis is increased for individuals with diabetes, research has also begun to show that periodontitis may contribute as a risk factor for diabetes onset and regulation. One study demonstrated that individuals with severe periodontitis were at increased risk for poor glycemic control upon followup [72]. To further investigate the effect that periodontitis has on the incidence of diabetes, investigators evaluated almost 3000 diabetes-free participants. Individuals with periodontitis at baseline demonstrated a greater increase in HbA1C values (a measure of diabetes) compared to those without a form of periodontitis at the 5-year follow-up. Thus, periodontitis predicted increases in HbA1C levels in diabetes-free individuals [73].

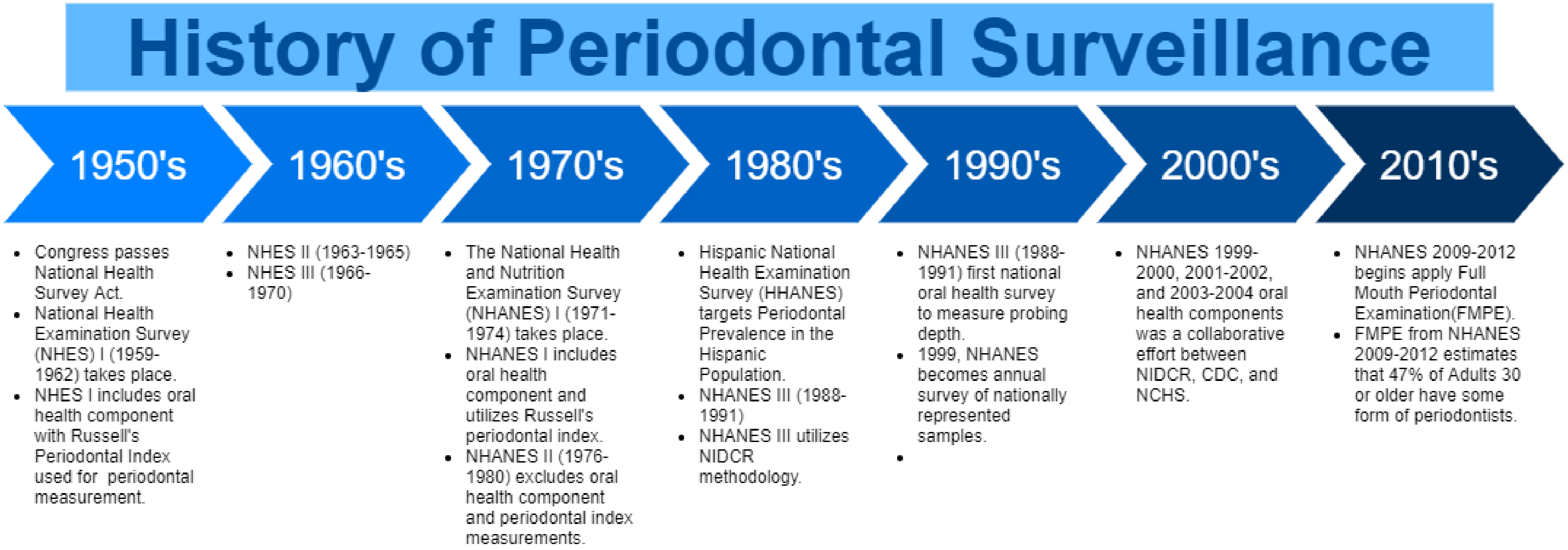
1.7 Systemic Diseases and Environment-Wide Association Studies (EWAS)

Previously, Genome-Wide Association Studies (GWAS) identified genetic markers associated with a myriad of disease types. However, the multi-faceted nature of systemic diseases, necessitates a broader viewpoint. Environment-Wide Association Studies (EWAS), when possible, can be used to highlight environmental factors associated with a disease. Patel et. al [74]utilized the NHANES dataset to conduct an EWAS to further understand the multi-faceted relationship between diabetes mellititus and various environmental variables captured by the survey. Because T2DM and periodontal disease are both systemic diseases, Dr.Li [75] took a similar approach as Patel for periodontal disease. This used the NHANES dataset but applied the EWAS approach to periodontal disease, expanding Patels methodological work by incorporating CART and Random Forests analyses in the investigation of the multi-faceted environmental factor-disease relationships. Chapter 2 expands the results of this dissertation providing novel insights for periodontal epidemiology.

1.8 History of Periodontal Surveillance within NHES and NHANES

In 1956, Congress enacted the National Health Survey Act to monitor the prevalence of all types of diseases across the U.S. and would later be the foundation of funding for the National Health Examination Survey (NHES). Later, a nutritional component was added, making the National Health and Nutrition Examination Survey (NHANES).

Figure 1.1: History Of Periodontal Surveillance



History of Periodontal Surveillance

1950's

- Congress passes National Health Survey Act.
- National Health Examination Survey (NHES) I (1959-1962) takes place.
- NHES I includes oral health component with Russell's Periodontal Index used for periodontal measurement.

1960's

- NHES II (1963-1965)
- NHES III (1966-1970)

1970's

- The National Health and Nutrition Examination Survey (NHANES) I (1971-1974) takes place.
- NHANES I includes oral health component and utilizes Russell's periodontal index.
- NHANES II (1976-1980) excludes oral health component and periodontal index measurements.

1980's

- Hispanic National Health Examination Survey (HHANES) targets Periodontal Prevalence in the Hispanic Population.
- NHANES III (1988-1991)
- NHANES III utilizes NIDCR methodology.

1990's

- NHANES III (1988-1991) first national oral health survey to measure probing depth.
- 1999, NHANES becomes annual survey of nationally represented samples.
-

2000's

- NHANES 1999-2000, 2001-2002, and 2003-2004 oral health components was a collaborative effort between NIDCR, CDC, and NCHS.

2010's

- NHANES 2009-2012 begins apply Full Mouth Periodontal Examination(FMPE).
- FMPE from NHANES 2009-2012 estimates that 47% of Adults 30 or older have some form of periodontists.

1.8.1 National Health Examination Survey (NHES)

1959-1970

NHES phases I through III included an oral health component using Russells periodontal index (PI) for measuring periodontal status of the population [76]. Before NHES IV was conducted, a nutritional component was added and NHES was renamed as the NHANES.

1.8.2 National Health and Nutrition Examination Survey (NHANES)

1971-1987

NHANES I took places from 1971-1974 and continued to utilize the PI method for periodontal status. NHANES II excluded the oral health component, as well as PI. Over time, inadequacies of the PI method were recognized as scores were weighted and based on a progression from gingivitis to periodontitis. Because the prevalence was difficult to ascertain within Hispanic populations in the United States from the national estimations, a separate survey for conducted in 1982 to 1984 targeting Hispanics [3]. After this survey, the PI method was overwhelmingly recognized as outdated and utilizing measures of clinical attachment became prominent and continued until 2005 [76]. In 1985-1986 the National Institute of Dental Research (now National Institute of Dental and Craniofacial Research; NIDCR) conducted a National Survey of Oral Health in the United States Employed Adults and Seniors [77]. For the first time, pocket probing depth was measured and loss of attachment was calculated. With this much information, varying case definitions could be developed from the same data. This was also the first time a partial-mouth periodontal examination (PMPE) was utilized reflecting cost and time constraints for conducting the epidemiologic study. During PMPE, only a limited number of teeth are examined for periodontal lesions, as well as a limited number of sites. While the information being collected from this survey was reflective of the etiology of the disease (not being a continuum from gingivitis to periodontitis), the PMPE method complicated the reporting of the disease as it could underestimate periodontal prevalence in populations [78].

1988-2004

NHANES III was a collaborative effort between the NIDCR and the National Center for Health Services (NCHS), and therefore, utilized the same methodologies from the previous NIDCR study. NHANES III was unique from previous cycles, as it would have 2 national study periods. Phase I took place from 1988 to 1991 and Phase II took place from 1988 to 1994 [79]. NHANES III was the first national oral health survey to measure probing depth, gingival recession and attachment loss in the United States [80] [81]. Previously, NHANES cycles had been completed periodically. However, in 1999, NHANES became an annual survey of nationally representative population samples [76]. The NHANES III periodontal protocol was continued in 1999 and 2000 with 2 periodontal sites being measured. Beginning in 2001, a third site was added [82]. The 1999 to 2004 NHANES oral health component was a collaborative effort between NIDCR, the CDCs Division of Oral Health, and NCHS.

2005-Present(2017)

In 2005, funding for the oral health component of NHANES decreased, directly affecting the periodontal component. From 2009-2012 the NHANES began applying a Full Mouth Periodontal Examination (FMPE) for survey participants [15].

1.9 Prevention

The prevention of periodontal disease can be seen in 3 stages: primary, secondary, and tertiary prevention. Primary prevention is aimed at preventing the inception of the disease. It includes strategies such as improving groups' and individuals' oral hygeine through education on protective strategies. The goal of secondary prevention is to impede the progression of periodontal disease severity at the earliest stage possible.

Tertiary prevention has the goal to improve the functional limitations as a result of the disease such as the restoration of missing teeth through implants and prosthetics. [83].

1.9.1 Treatment Methods

In 2017, a review was released in *Periodontology 2000* that examined the current status of periodontology as well as the challenges that remain for the future. This review also highlighted current treatment methodologies and their efficacy. [84]

Since the prevention (primary prevention) and inhibition of the progression of periodontal disease (secondary prevention) both involve limiting biofilm accumulation, it stands to reason that patient self-care is an area of research that demands attention. Self-care includes oral hygeine instruction, tooth-brushing, the use of dentrifices (toothpowders, toothpaste, and gels) in combination with a tooth-brush, interdental devices, dental floss, and oral irrigators. Oral hygiene products are usually tested less than 12 months time in research, and have yet to prove their efficacy in preventing toothlessness via periodontal disease in a lifetime. Moreover, these products have been shown to be effective in mild forms of periodontal disease such as gingivitis and mild periodontitis which are not representative of all the adult forms of periodontal disease.

This review differentiated the treatment efficacies based on severity. Mild and moderate periodontitis should be treated using non-surgical therapies and contain an educational component on self-care. However, severe periodontitis does not respond to these tradition methodologies. The current protocol for severe periodontitis is some form of surgical treatment, but the research is lacking for it to be ”best practices” for the standard of care. Antibiotics have demonstrated similar results as surgical intervention. Periodontal surgery lacks the controlled clinical trials to back its widespread application on patients.

In terms of the etiology, a dysregulated immune system is a key component to the development of periodontal disease [85]. It has been suggested that more severe forms of periodontal disease could utilize immunotherapy as an adjunct or alternative to traditional therapy. If clinicians were to utilize immunosuppressors across the majority of periodontal cases (most of which would include chronic periodontitis), this might compromise the protective nature of the immune system and lead to further complications.

1.9.2 Public Health Approaches

A variety of approaches have been suggested to improve prevention of periodontal disease among populations. Public health practicioners suggest affecting health behaviors (smoking, oral health practices), as well as targeting high risk sub-populations for educational initiatives. Currently, dental patients are encouraged and taught about proper oral health behaviors and given "chair-side" advice on smoking behaviors [86]. The effects of non-consistent educational approaches has been shown to be limited and insufficient to affect the prevalence in a population. The high cost for periodontal treatment and necessary follow-ups precludes proper treatment for the economically disadvantaged groups. In addition, periodontal treatment is not always covered by private dental plans or federal and state funded programs such as medicaid or medicare. According to Medicaid and CHIP Payment and Access Commission, as of February 2015 Medicaid programs are required to cover dental services for children and youth under age 21. As a part of Medicaid, adult periodontal services were covered in 19 states as of February 2015 [87].

Interprofessional approaches is now gaining greater attention for its potential to effect multiple diseases. This is best summarized by the following article segment from the project, "Advancing Dental Education in the 21st Century."

"Reducing the risk of disease can be accomplished by an emphasis on smoking cessation and dietary intake and the prevention of obesity. These activities will promote interprofessional health care education and practice. While change is always challenging, this new practice paradigm could improve both oral health and health outcomes of patients seen in the dental office." [88].

Chapter 2 The Exposome and Periodontal Disease: Epidemiologic Evaluation of NHANES within Smoking Classification

2.1 Introduction

This chapter is a direct insertion of a paper that is the result of joint work with myself, Dr. Jeffrey Ebersole, Dr. Pinar Emecen Huja, Dr. Grace Li, and Dr. Heather Bush. Specifically, I assisted in the gathering of the datasets from NHANES, programming SAS and R code, running the CART and Random Forest Models, and assisting in writing of the paper and providing figures and tables. This paper will be submitted in the Fall of 2017.

The Exposome and Periodontal Disease: Epidemiologic Evaluation of NHANES within Smoking Classification

P. Emecen-Huja[1], H. Li[2], J.L. Ebersole[1,3], J. Lambert[2], and H. Bush[2]

[1]Division of Periodontics, College of Medicine, [2]Department of Biostatistics, College of Public Health, and [3]Center for Oral Health Research, College of Dentistry, University of Kentucky, Lexington, KY

Running Title: Environment and periodontitis

Corresponding Author: Dr. Pinar Emecen-Huja

Keywords: environment, periodontal disease, NHANES

Summary

Periodontitis is an inflammation of the gingival tissues caused by the accumulation of bacterial biofilms that can be can be affected by environmental factors. **Objective:** This report describes an Environment Wide Association Study (EWAS) to evaluate the relationship of the exposome to the expression of periodontitis using the National Health and Nutrition Examination Study (NHANES) from 1999-2004. **Material and Methods:** Environmental variables (156) were assessed in patients categorized for periodontitis (n=8884). Multiple statistical approaches were used to explore this dataset and identify environmental variable patterns that significantly enhanced or lowered the prevalence of periodontitis. **Results:** An array of environmental variables were significantly different in periodontitis in smokers, former smokers, or non-smokers, with a subset of specific environmental variables identified in each population subset. Discriminating environmental factors included blood levels of lead, phthalates, selected nutrients, and PCBs. Importantly, these factors were found to be coupled with more classical risk factors (ie. age, gender, race/ethnicity) to create a model that predicted an increased disease likelihood of 2-4 fold across the population. **Conclusions:** Targeted environmental factors are significantly associated with the prevalence of periodontitis. Existing evidence suggests that these may contribute to altered gene expression and biologic processes that enhance inflammatory tissue destruction.

Clinical Relevance

Scientific rationale for the study: This report describes the use of an Environmental Wide Association Study (EWAS) of NHANES data to identify the contribution of environmental risk factors to periodontal disease prevalence.

Principal findings: Our findings indicate a subset of environmental factors combined with classical risk factors for periodontal disease, age, gender and race, enhanced the prevalence of periodontal disease 2-4 fold.

Practical implications: Environmental risk factors, alone or in combination with genetic and epigenetic factors, may be used for early risk profiling of clinical expression of periodontal disease.

INTRODUCTION

Despite increasing awareness and improvement in oral health, periodontitis, together with dental caries, remain major health concerns across the lifespan in the United States (Benjamin, 2010). Periodontal disease occurs as a result of an interaction between bacterial biofilms and immunoinflammatory responses. It is anticipated that 80% of the risk for periodontal tissue damage is a result of dysregulated host responses against the chronic bacterial insult (Grossi et al., 1994, Roberts and Darveau, 2015, Pihlstrom et al., 2005). This interaction can progress to destroy the periodontal tissues and bone, and eventually is the major basis of tooth loss in adults with edentulous individuals having difficulty eating, swallowing, and speaking properly (Petersen and Ogawa, 2005, Bartold et al., 2010, Darveau, 2010). These impaired oral functions can greatly impact individual quality of life, negatively affecting societal and economic opportunities, and continues to expand as a public health concern in aging populations (Jansson et al., 2014).

Similar to many chronic diseases, it is well documented that periodontal disease is a complex disease with multiple potential contributing factors. These include genetic and epigenetic influences, patient behaviors, medication use, and/or environmental factors, which all together promote periodontal disease initiation and progression (Meyle and Chapple, 2015). Low socioeconomic status, poor oral hygiene, psychological stress, depression, increased age, Hispanic ethnicity, diet/obesity, and systemic health co-morbidities are well known risk factors that contribute to the prevalence of periodontal diseases (Albandar, 2002, Albandar, 2005, Stabholz et al., 2010). However, smoking has been identified as one of the most significant and modifiable risk factor in the pathogenesis of periodontitis and tooth loss (Bergstrom and Preber, 1994, Dietrich et al., 2015). Data also support that the number of cigarettes smoked per day is directly related to the prevalence and the severity of the disease (Eke et al., 2016, Tomar and Asma, 2000, Martinez-Canut P, 2005). Emphasis has been placed on the need for more effective management of these modifiable risk factors to impact this global disease (Van Dyke and Sheilesh, 2005), albeit, non-modifiable factors including age, genetics and the existence of various systemic diseases are clearly more challenging to address across the population (Van Dyke and Sheilesh, 2005, Loos et al., 2015, Kaye et al., 2016, Reynolds, 2014).

In this regard, an array of studies of this complex disease have provided evidence attributing disease expression and severity to genetic predisposition for the characteristics of the host response to the oral microbial challenge. These have included genes controlling the production of inflammatory mediators and tissue and bone regulatory molecules (Chantarangsu et al., 2016, Lavu et al., 2015, Wu et al., 2015, Ding et al., 2014, Scapoli et al., 2015). A number of reports describing genetic modifications within affected families (Michalowicz et al., 2000) and more recently using Genome Wide Association Studies (Feng et al., 2014, Rhodin et al., 2014) have identified genetic variations and polymorphisms that associate with the expression of periodontal disease. More recently, studies reported epigenetic alterations in the genomes of periodontitis patients that may provide some additional mechanistic explanations to variation in patient clinical responses beyond only the gene sequences (*ie.* single nucleotide polymorphisms) (Barros and Offenbacher, 2014, Larsson et al., 2012, Loo et al., 2010). Importantly, studies from other disease models show that various environmental stimuli can contribute to these epigenetic changes and underpin the concept of environment-gene interactions related to disease expression (Vaiserman, 2014). While rather limited data is available regarding environmental factors in periodontitis (Saraiva et al., 2007), the National Health and Nutrition Examination Survey (NHANES) provides a robust data set regarding measures of 156 environmental factors in blood and urine. This report describes the use of various epidemiologic and statistical tools to conduct an Environment-Wide Association Study [EWAS; (Patel and Manrai, 2015)] with periodontitis in the U.S. adult population.

MATERIALS and METHODS

Population data:

In this study, periodontal examination data from three NHANES cohorts, 1999-2000, 2001-2002, 2003-2004, were extracted and combined to comprise the study population. Among the 11,837 participants who were equal to and older than 18 years of age and had at least 16 teeth, 3,745 were collected in the first cohort (1999-2000), 4,258 in the second cohort (2001-2002), and 3,834 in the third cohort (2003-2004). Those with missing smoking status and periodontal parameters were excluded leaving a final analytical sample of 8,884 participants.

Demographics:

The demographic variables considered in this study included age, gender, race, socio-economic status, smoking status, and number of teeth. Racial-ethnic groups were summarized into five categories: Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, and Other Race. Socio-economic status, estimated using the poverty income ratio, was computed as the ratio of family/individual income to the appropriate federal poverty threshold. Smoking status, current smoker, former smoker, non-smoker, was derived from the two self-reported questions. Participants reported having historically smoked more than 100 cigarettes, but currently not smoking were defined as former smokers. Non-smokers were defined as reporting never smoking.

Clinical parameters:

Periodontitis was defined as a site with clinical attachment loss (CAL) ≥3 mm and a periodontal pocket ≥4 mm. NHANES (1999-2004) uses the partial-mouth periodontal examination (PMPE) protocol to sample teeth and sites. The PMPE protocols randomly selects two quadrants of the mouth and specified 2 to 3 sites per tooth for measurement of pocket depth, attachment loss, and bleed on probing. In 1999-2000, two sites per tooth (mid-facial and mesio-facial) were measured, while three sites per tooth (mid-facial, mesio-facial and distal) were measured in 2001-2002 and 2003-2004.

Environmental variables:

The environmental factors were categorized into 15 classes based on NHANES categorization. Environmental variables measured in at least one of the three data cohorts (i.e. 1999-2004) were included in the study. A total of 156 environmental factors were measured in the NHANES data using blood and urine samples. These included chemical toxicants, pollutants, allergens, bacterial/viral organisms and nutrients. Environmental factors with laboratory measurements that had less than 90% of the observations below a detection limit threshold defined by NHANES were omitted from analysis. Since most of the variables were measured by mass spectrometry and absorption spectroscopy the laboratory measurements were detected in small ranges and were skewed. These variables were log-transformed (natural) and standardized and referred to as “processed”.

Statistical approaches:

Survey-weighted logistic regressions were performed for each of the processed environmental factors, adjusting for age, gender, ethnicity, socio-economic status, smoking status and number of teeth. The R package “survey” was used in R (Version 3.1.2) for the survey-weighted logistic regression. Weights were constructed in SAS (Version 9.4) using a 6 year weighting design from the NHANES variable WTMEC2YR73 (http://www.cdc.gov/nchs/tutorials/Nhanes/SurveyDesign/Weighting/Task2.htm). Adjusted odds ratios were calculated with 95% confidence intervals were provided to demonstrate the relationship between the individual factors and periodontitis. These regressions were repeated by smoking status to examine potential relationships within smoking categories. We used a significance level of p=0.01 to select significant associations

This Environmental Wide Association Study (EWAS) analysis employed random forests (RF) and classification and regression trees (CART) to investigate associations and potential interactions between environmental factors, demographic and socioeconomic characteristics, and periodontitis disease status (Breiman, 2001). Specifically, RF was used to identify important factors (main effects and interactions) and CART was used to visually investigate these relationships. These methods were selected because the data involved many potentially correlated environmental factors and had the ability to allow nonlinearities and interactions without modeling them explicitly (Strobl et al., 2009). These analyses were performed using the “party”(Version 1.0-25) package in R (Version 3.1.2).

RESULTS

The final statistical analysis was completed on 8,884 individual who were $\geq$18 years old and had at least 16 teeth. Males comprised 48.4% of the sample. The majority of subjects were non-smokers (55.9%), and those with smoking experience were evenly distributed between former smokers (22.5%) and current smokers (21.6%). The ethnic distribution of the group was non-Hispanic white (48.5%), non-Hispanic black (18%), Mexican American (25.1%), other Hispanic (4.7%), and other race including multi-racial (3.7%). Approximately 72% of the sample population was older than 30 years of age. (**Table 2.1 & 2.2**). The weighted prevalence of periodontitis was 8.1% across the entire >18 years of age population. When the periodontitis group was compared to the non-periodontitis subset, individuals with periodontal disease were more likely to be male, older than 30 years of age, Mexican American, non-Hispanic black or Hispanic compared to Non-Hispanic white and current smoker (p<0.001) (**Table 2.1 & 2.3)**.

Using survey-weighted logistic regression, there were 42 environmental factors (cotinine, 1 dioxin, 4 heavy metals, 8 hydrocarbons, 8 nutrients, 17 PCBs and 3 volatile compounds) that resulted in a statistically significant adjusted odds ratio for disease versus health in this NHANES cohort (**Table 2.4**). When data were stratified by the smoking status 19 environmental factors (2 dioxins 2 heavy metals, 2 nutrients, 1 phthalate, and 12 PCBs) in current smokers, 13 factors (acrylamide, 1 heavy metal, 1 nutrient, 1 phthalate and 9 PCBs) in former smokers, and 12 factors (1 dioxin, 2 heavy metals, 2 nutrients, 1 pesticide, and 6 PCBs) in non-smokers were identified (**Table 2.5**).

In regression analyses considering each environmental factor separately, blood lead levels were consistently identified as a significant factor in both the overall and stratified analyses ([a]OR =1.54 for current smokers; [a]OR=1.39 for non-smokers; [a]OR=1.57 for former smokers) (**Table 2.5**). Among the 17 polychlorinated biphenyls (PCBs) found to be associated with periodontitis in the overall sample, 6 (i.e. PCB105, PCB157, PCB172, PCB177, PCB178, and

PCB206) were also found significantly elevated in all patient subsets, with adjusted odds ratios ranging from 1.41 to 5.29. Also, across these environmental variables, the adjusted OR estimates were significantly lowerin non-smokers compared to current and former smokers. The smoking population also demonstrated additional factors, including 6 PCBs (PCB66, PCB146, PCB167, PCD170, PCB183, PCB187) with adjusted OR estimates from 1.63-2.23, dioxins (PNCDD, TCDD) with adjusted OR estimates of 1.66 and 1.81, and blood nutrients retinyl stearate and retinyl palmitate with adjusted OR estimates from 1.32-1.35.   In contrast, blood nutrients such as Vitamin D and cis-ß-carotene were found to be significantly protective for periodontitis.Higher levels of Vitamin D were estimated to decreasing the odds of periodontitis by 39% and 24% in former and non-smoker groups, respectively ([a]OR=0.61 for former smokers; [a]OR=0.76 for non-smokers), and cis-ß-carotene estimated at decreasing the odds for periodontitis by 22% in non-smokers ([a]OR=0.78) (**Table 2.5**). Utilization of a Random Forests analysis was also used to identify environmental variables that were related to periodontitis.

We subsequently employed a Classification and Regression Tree (CART) analysis to identify and visualize relationships of critical demographic and environmental factors. Based upon the variation in factors that were clearly related to smoking, the CART was performed separately for each of the smoking, former smoking, and non-smoking subsets. **Figure 2.1** provides a depiction of the CART for each of these cohorts. Within the smoking group, **Figure 2.1A** presents a CART analysis of the variables that demonstrated elevated blood lead levels as an initial discriminator, with age >31 yrs. stratifying patients with an approximate 4-fold risk of periodontitis. **Figure 2.1B** visualizes the factors classifying the disease risk in former smokers. In this case race/ethnicity remained a critical factor. Those who reported race/ethnicity other than non-Hispanic white demonstrated increased disease prevalence; elevated blood lead levels and age >53 increased periodontitis prevalence to 37%. Within the subset of non-Hispanic white subjects and other race including multi-racial, a prevalence rate of 12% was observed in those with elevated blood lead levels.

For non-smokers, which comprised 56% of the total population, multiple variables were identified to have relationships with periodontal disease status. Race-ethnicity and age were important distinguishing factors. Prevalence rates were low across those reporting non-Hispanic white race, but even in this group subjects >57 years with higher blood lead levels demonstrated an increased prevalence. The prevalence was further modified by elevated urine antimony that increased the observed prevalence of periodontitis to 33% from as low as 8% in the low urine antimony and high cis-ß-carotene group. (**Figure 2.1C**). For those with low levels of cis-ß-carotene, higher blood lead levels showed a higher prevalence of periodontitis of 18% compared to 11% for the lower blood lead group

DISCUSSION

It is clear that periodontitis represents a dysregulation of the host response to a dysbiotic microbiome that occurs in a large portion of the global population. Substantial work is being conducted via the Human Microbiome Project (Cross et al., 2016) to discern not only the characteristics of the alterations in the disease microbiome, but also interrogating complex metagenomic datasets to assess functional changes in the microbial ecology associated with health and disease (Madupu et al., 2013). Additionally, a complementary research direction is attempting to document the role of individual genetic variation across the population that contributes to disease expression and severity (Loos et al., 2015). These studies have employed

SNP analysis of specific targeted genes (Scapoli et al., 2015, Wu et al., 2015, Loos et al., 2015), Genome-Wide Association Studies (GWAS) (Divaris et al., 2013, Rhodin et al., 2014) and epigenetic analyses (Larsson et al., 2015, Barros and Offenbacher, 2014, Zhang et al., 2010) to help elucidate the complex of factors that interact to create a disease susceptible host. This report describes an additional consideration in disease expression focused on the larger environmental variation to which individual members and subgroups of the U.S. population are exposed (*ie.* exposome) as a potential direct contributor to the microbial dysbiosis (Hajishengallis, 2014) and/or a modifier of host responses through altered molecular pathways or modulation of genetic control of the disease (Larsson et al., 2012, Schulz et al., 2016). The findings identified more classical factors (*ie.* age, gender, race/ethnicity) in the disease model, but for the first time integrated a subset of environmental factors, both toxins and nutrients, that appear to substantially modify the relative risk for periodontitis in the population. The identification of the association of environmental toxins including lead, hydrocarbons, polychlorinated biphenyls, and nutrients such as retinyl stearate in models described a significant increase in relative risk of disease. Thus, the findings support the potential for a role of these factors in modifying the challenge (ie. bacterial biofilms) and/or host responses with a loss of homeostasis and tissue destruction.

The results demonstrated altered levels of various heavy metals, including lead, cadmium and antimony in periodontitis patients. A range of literature has shown the toxic properties of systemic elevations in heavy metals from environmental sources, including lead (Dahl et al., 2014, Lucchini and Hashim, 2015). In particular, this toxin has been linked to substantial neurotoxicity and negative developmental processes in children (Zhang et al., 2013, Senut et al., 2012). This study identified, using CART analysis, a threshold of >2.0 µg/dL that discriminated periodontitis from health in the adult population. While this level does not indicate the actual blood lead level across the periodontitis group, since CART attempts to fit the discrimination profile in the context of multiple variables, it was clear that in all subsets of smokers, former smokers and non-smokers that lead levels are significantly elevated in periodontitis patients. An earlier evaluation of data from NHANES III (1988-94) demonstrated a significantly increased OR for periodontitis in both men and women with increased blood lead levels (Saraiva et al., 2007). Reports examining various iterations of the Korean NHANES (KHANES) study demonstrated elevated lead, cadmium, or mercury in subjects with periodontitis, particularly related to smoking and in some instances gender associated similar to our data from NHANES (Won et al., 2013, Kim and Lee, 2013, Rhee et al., 2013, Moon, 2013). An additional study reported that chronic occupational exposure of workers to lead resulted in significant changes in oral health and correlated with increasing blood lead levels (El-Said KF, 2008). Terrizzi et al., have reported that elevated lead levels under hypoxia induces alveolar bone resorption and periodontitis (Terrizzi et al., 2013). More recently they demonstrated that iNOS and PGE_2 levels are altered by lead and hypoxia as inflammatory responses that would contribute to damage of the periodontium (Terrizzi et al., 2014). Additional studies have also indicated that elevated lead levels within the environment of various cell types will alter their functions, albeit, the majority of these *in vitro* studies have focused on neuronal cell targets (Song et al., 2016, Hu et al., 2011). Moreover, the lead levels investigated in these previous studies generally targeted levels that have been shown in blood to have substantial neurotoxicity ($\geq$10 µg/dL), although levels of >5 µg/dL are considered deleterious (Rahman et al., 2011). Further studies will be required to identify the relationship of blood lead levels to severity of the disease, age of onset, and response to therapy, as well as biologic studies determining the impact of these altered levels of lead on host responses, and even the microbial ecology related to the disease process.

Polychlorinated biphenyls (PCBs) were once widely deployed as dielectric and coolant fluids in electrical apparatus, carbonless copy paper, and in heat transfer fluids since they do not easily degrade. PCBs' environmental toxicity and classification as a persistent organic pollutant resulting in production and use of them being banned by the United States Congress in 1979. Coplanar PCBs, *eg.* dioxin-like PCBs, since their structure is similar to dioxins, allows them to act as agonists of the aryl hydrocarbon receptor (AhR). They are considered as contributors to overall dioxin toxicity within the environment. The toxicity of PCBs varies considerably among various chemical structural iterations with the coplanar PCBs representing 12/209 possible PCB molecules (*ie.* PCB 77, 81, 114, 118, 123, 126, 156, 157, 167, 169, 189) generally considered among the most toxic congeners with the majority of differences occurring in smokers and former smokers. Interestingly, the overall group of toxins included PCB105, PCB146, PCB172, PCB177, PCB178, PCB183, and PCB206, which are all members of the non-coplanar group of PCBs appeared to show the most frequent association with periodontitis. Non-coplanar PCBs cause neurotoxic and immunotoxic effects, but at levels much higher than normally associated with the dioxins congeners and do not activate the AhR (Levin et al., 2005, Hamers et al., 2011). The non-coplanar PCBs have been suggested to function by interfering with intracellular signal transduction and critical calcium transport mechanisms (Mundy et al., 1999, Tilson, 1998). The PCBs readily penetrate skin and other epithelial barriers and as fat-soluble compounds they can be at 100-200 times higher levels in adipose tissue than in serum (Grimm et al., 2015). Elevated levels of non-coplanar PCBs, including PCB153, PCB170, PCB180 and PCB187 were detected in the blood of Canadian First Nations communities and were associated with elevated levels of an array of immune activation markers including IFNγ, IL-1ß, IL-8, IL-17A and TNFα (Imbeault et al., 2012). Much of the molecular aspects of PCBs and host responses have focused on the coplanar, dioxin like congeners.

Coplanar-type congener PCBs demonstrate that these AhR ligands induce oxidative stress and increased translocation of NFκB in endothelial cells, as well as significantly altering inflammatory/immune responses in adipocytes potentially via upregulation of expression of AhR on the cells (Kim et al., 2012). This altered pathway activation increased expression of IL-6 and VCAM-1 as markers of inflammatory reactions (Hennig et al., 2002). PCB126 exposure increased endothelial cell inflammatory responses including CRP, IL-6 and IL-1ß (Liu et al., 2016), as well as significantly altering cellular adhesion molecules, ICAM-1 and VCAM-1, that would impact normal vascular functions (Kumar et al., 2014). PCBs have been proposed to alter host responses via histone modifications and epigenetic changes. Coplanar PCB77 and PCB126 also altered an array of vascular cell inflammatory mediators via altered functions of NFκB (Liu et al., 2015). Additionally, coplanar PCBs upregulated MCP-1 production by endothelial cells, *ie.* proinflammatory, that was modulated by treatment with the omega-3 fatty acids, EPA or DHA (Majkova et al., 2011). *In vivo*, various types of PCBs (PCB77, 104 and 153) were administered orally to mice and significantly increased the expression of proinflammatory biomolecules (Sipka et al., 2008). Finally, a single recent report demonstrated that PCB126 appeared to exacerbate periodontal disease in a susceptible species of mink (Ellick et al., 2013). The current study identified an array of PCBs that were increased across the periodontitis population. While some representative true dioxin molecules significantly increased the OR for periodontitis these only were noted in smokers. No other reports are available identifying PCB levels and periodontitis in humans or animal models, nor focusing on biologic alterations in cells related to periodontal health and disease, thus, this family of exposome factors could present an important area for further investigation of disease variation and personalized documentation of disease features within the population.

An interesting finding was the dichotomy between the effects of selected specific nutrients on the expression of periodontitis. Both carotenoids and Vitamin D levels showed significant Odds Ratios for protecting against periodontitis. Carotenoids are organic pigments found in plants and some photosynthetic microorganisms and carotenoids from human diets are stored in the fatty tissues. There are over 600 known carotenoids classified as xanthophylls (β-cryptoxanthin, lutei, and zeaxanthin; non-vitamin A carotenoids) and carotenes (α-carotene, β-carotene, and lycopene). Generally, the health benefits of carotenoids are thought to be due to their role as antioxidants with dietary carotenoids proposed to interact with endogenous antioxidant enzymes to positively affect immunity (Babin et al., 2015). Thus, various reports have shown that elevations in acute phase proteins are accompanied by low vitamin A levels (Thurnham et al., 2015) and that carotenoids significantly reduced proinflammatory cytokines, CRP, and other markers of inflammation in multiple tissues (Gammone et al., 2015). A study of inflammation in 60-70 year old men demonstrated an inverse relationship between elevated carotenoids and serum CRP levels.(Cao et al., 2016) Rodent models of carotenoid administration have shown lowered oxidative stress and nitric oxide synthase and associated inflammation in rats (Xu et al., 2015) and substantial anti-inflammatory effects on CRP, TNFα, IL-1ß and IL-6 in mice following an LPS-induced inflammatory challenge (Firdous et al., 2015). Molecularly, carotenoids can impact intracellular signaling pathways, such as NFκB thus influencing gene expression patterns and inflammatory mediator profiles (Kaulmann and Bohn, 2014). These effects have been evaluated in various studies related to periodontal disease. ß-cryptoxanthin suppressed LPS-induced osteoclast formation and lowered alveolar bone loos in a mouse model of disease (Matsumoto et al., 2013) and decreased *P. gingivalis*-induced IL-6 and IL-8 production by human periodontal ligament cells (Nishigaki et al., 2013). Moreover, low blood levels of various carotenoids have been associated with an increased prevalence of periodontitis in 60-70 year old men (Linden et al., 2009) and carotenoid levels were related to positive outcomes of scaling and root planning with the relationship limited to non-smokers (Dodington et al., 2015). Thus, our data from a large population cohort is consistent with these findings and the support that increased availability of carotenoids appears to provide some level of protection from periodontitis.

Vitamin D has received an increasingly detailed examination regarding its potential influence in periodontitis. Various reports have linked decreased serum or saliva vitamin D levels with tooth loss and periodontitis (Zhan 2014; Abreu 2016, Joseph 2015; Antonoglou 2015; Gumus 2016) including in smokers (Lee 2015), albeit not all studies are supportive since this was not observed in postmenopausal women (Pavlesen 2016). Additionally, a gene polymorphism for vitamin D binding protein increases the risk for periodontitis (Song 2016) that appears exacerbated in smokers (Chantarangsu 2016). Mechanistically, vitamin D has been shown to alter virulence gene expression in *P. gingivalis* and downregulate NFκB activation and cytokine secretion by monotyes and macrophages (Grenier 2016; Xu 2016). Moreover, vitamin D has been reported to improve innate immune functions and antimicrobial peptide production by epithelial cells at mucosal surfaces (McMahon 2011; Rigo 2012; Dhawan 2015). Our epidemiologic analysis of this nutrient was based upon examination of NHANES data demonstrating a significant protective feature of this serum nutrient in periodontitis, specifically in non-smokers and former smokers. This type of finding is consistent with additional associational data from NHANES related to risk of cardiometabolic disease (Al-Khalidi 2017), asthma (Han 2016), and coronary heart disease and all-cause mortality (Daraghmeh 2016). Interestingly, a single recent report describes the interaction of an environmental exposure to phthalates may decrease blood levels of vitamin D (Johns 2016), an observation consistent with our results identifying “competing” impact of environmental toxins and nutrients on periodontitis as the clinical outcome.

In contrast, elevated levels of retinyl stearate and retinyl palmitate each significantly enhanced the risk for periodontitis particularly in smokers. The retinoids comprise a class of compounds related to Vitamin A. These compounds have been used to regulate epithelial cell growth, as well as playing a role in vision, regulation of cell proliferation and differentiation, growth of bone tissue, immune functions, and even activation of tumor suppressor genes (Reichrath et al., 2007). Our data demonstrated a significantly increased OR for blood levels of retinyl stearate and retinyl palmitate in periodontitis. In serum, 56% of retinyl esters is retinyl stearate, 33% retinyl palmitate, and 5% retinyl oleate. Retinyl esters in humans are derived from animal sources and are hydrolyzed in the intestinal lumen to form retinol and fatty acids, such as retinyl palmitate or stearate. Enzymes in the intestinal lumen that hydrolyze dietary retinyl esters include cholesterol esterase from the pancreas and a retinyl ester hydrolase intrinsic to cells of the small intestine, which primarily acts on long-chain fatty acids, such as palmitate or stearate

(Reichrath et al., 2007). Generally, retinol is taken up by the absorptive cells of the small intestine in the all-trans-retinol form. Biologically, altered vitamin A levels are associated with dysregulation of cytokine/chemokine production that impact immunity and inflammation (Spinas et al., 2015). Retinoic acid has been shown to increase proinflammatory mediators in skin mast cells (Babina et al., 2015). Retinyl palmitate was also identified to counteract oxidative stress reactions in an animal model of septic shock (Basu, 2001). A single study has been reported regarding these compounds and periodontitis. Wang et al. (Wang et al., 2014) demonstrated that all-trans retinoic acid administration modulated the Th17/Treg balance and can modulate the expression of periodontitis in a murine model of *P. gingivalis* infection and provided protection against periodontitis with increased Treg activation and decreased Th17 functions. However, our data specifically related to endogenous levels of a specific retinoid, retinyl stearate, suggested an increased risk for periodontitis. This may relate to the more individualized functions of the various members of this family of dietary nutrients, and may highlight some unique features of the diet or intrinsic variation in the hydrolytic enzymes across the population that may link retinyl stearate and disease. Clearly additional studies will need to be conducted examining in more detail the clinical relationship with this compound, as well as its potential role in affecting an array of inflammatory responses that would be related to periodontitis.

Significantly elevated acrylamide levels were identified in the blood of former smokers, with an increased OR in the blood from smokers, as well. The majority of acrylamide is used to manufacture various polymers including those used in water treatment, as well as a binding and thickening agent in grout, cement, cosmetics, food packaging, plastic products, and paper production. Interestingly, acrylamide was discovered in prepared foods in 2002, specifically related to starchy foods, *eg.* potato chips, French fries, that had been heated higher than 120°C. (Tareke et al., 2002) Importantly, acrylamide is considered a potential occupational carcinogen by U.S. government agencies and cigarette smoking is a major acrylamide source in the population (Vesper et al., 2007) increasing blood levels by 3-fold versus other environmental sources. Recent findings demonstrate that chronic exposure to dietary acrylamide increases the rate of endothelial senescence that could impact the normal functional capabilities of the oral mucosal tissues (Sellier, 2015). Negligible information is available concerning the levels of this toxic substance and inflammatory/immune responses at mucosal surfaces.

Diethylphosphate is related to the family of alkyl phosphates termed organophosphates, which are widely distributed in nature related to high energy metabolites (eg. ATP), nucleic acids including both DNA and RNA, and are even the anti-HIV drug AZT that is active as an alkyl phosphate *in vivo*. However, the dialkyl phosphates are also metabolites of organophosphorous pesticides and represent human exposure to these pesticides within the environment. An early study using NHANES data examined an array of dialkyl phosphate metabolites from organophosphorous pesticides in the U.S. population (Barr 2004). They identified higher levels of these environmental toxins in young children that was unaffected by sex or racial/ethnic group. Minimal other information is available relating these compounds to diseases of mucosal surfaces. However, these pesticide metabolites have been shown to increase asthma-related cytokine levels in serum, particularly related to non-Th2 responses (Mwanga 2016). Related toxins have also been demonstrated to induce NLRP3 inflammasome activation and pyroptosis/apoptosis of keratinocytes via increased oxidative stress (Jang 2015). Thus, a combination of various environmental stressors altering fundamental balances in host cell functions that may already be stressed by a dysbiotic microbiome presents a new consideration regarding environment-gene interactions and expression of periodontitis.

This report describes an associational study of a large U.S. population sampled over an interval of 5 years via the NHANES project and demonstrated significant relationships of a subset of exposome challenges to the expression of periodontitis. A clear limitation in the approach is that the findings do not deliver any cause and effect relationship, and are affected by the lack of detailed clinical evaluation of periodontitis that is generally accepted within the field. However, the model developed identified an interaction of these exposome factors and more classical risk factors of age, gender, and race/ethnicity, thus providing some confidence that the findings are providing additional clues into population variation in disease expression. The model will also enable future testing with additional NHANES datasets, as well as the exposome features and categorization of disease. The individual exposome components that were identified can be further evaluated in more detailed clinical studies, and by implementing basic biologic studies of the host cells and microbiome components associated with health and disease to delineate modes of actions of these environmental factors that could contribute to the disease processes.

ACKNOWLEDGEMENTS

This work was supported by funding from the Center for Oral Health Research in the University of Kentucky College of Dentistry and from P20GM103538 and UL1TR000117 from the National Institute of Health.

REFERENCES

Albandar, J. M. (2002) Global risk factors and risk indicators for periodontal diseases. *Periodontol 2000* **29,** 177-206.

Albandar, J. M. (2005) Epidemiology and risk factors of periodontal diseases. *Dent Clin North Am* **49,** 517-532, v-vi. doi:10.1016/j.cden.2005.03.003.

Babin, A., Saciat, C., Teixeira, M., Troussard, J. P., Motreuil, S., Moreau, J. & Moret, Y. (2015) Limiting immunopathology: Interaction between carotenoids and enzymatic antioxidant defences. *Dev Comp Immunol* **49,** 278-281. doi:10.1016/j.dci.2014.12.007.

Babina, M., Guhl, S., Motakis, E., Artuc, M., Hazzan, T., Worm, M., Forrest, A. R. & Zuberbier, T. (2015) Retinoic acid potentiates inflammatory cytokines in human mast cells: identification of mast cells as prominent constituents of the skin retinoid network. *Mol Cell Endocrinol* **406,** 49-59. doi:10.1016/j.mce.2015.02.019.

Barros, S. P. & Offenbacher, S. (2014) Modifiable risk factors in periodontal disease: Epigenetic regulation of gene expression in the inflammatory response. *Periodontol 2000* **64,** 95-110. doi:10.1111/prd.12000.

Bartold, P. M., Cantley, M. D. & Haynes, D. R. (2010) Mechanisms and control of pathologic bone loss in periodontitis. *Periodontol 2000* **53,** 55-69. doi:10.1111/j.1600-0757.2010.00347.x.

Benjamin, R. M. (2010) Oral health: the silent epidemic. *Public Health Rep* **125,** 158-159.

Bergstrom, J. & Preber, H. (1994) Tobacco use as a risk factor. *J Periodontol* **65,** 545-550. doi:10.1902/jop.1994.65.5s.545.

Breiman, L. (2001) Random Forests. *Machine Learning* **45,** 5-32.

Cao, Y., Wittert, G., Taylor, A. W., Adams, R., Appleton, S. & Shi, Z. (2016) Nutrient patterns and chronic inflammation in a cohort of community dwelling middle-aged men. *Clin Nutr*. doi:10.1016/j.clnu.2016.06.018.

Chantarangsu, S., Sura, T., Mongkornkarn, S., Donsakul, K. & Torrungruang, K. (2016) Vitamin D Receptor Gene Polymorphism and Smoking in the Risk of Chronic Periodontitis. *J Periodontol* **87,** 1343-1351. doi:10.1902/jop.2016.160222.

Cross, B., Faustoferri, R. C. & Quivey, R. G., Jr. (2016) What are We Learning and What Can We Learn from the Human Oral Microbiome Project? *Curr Oral Health Rep* **3,** 56-63. doi:10.1007/s40496-016-0080-4.

Dahl, C., Sogaard, A. J., Tell, G. S., Flaten, T. P., Hongve, D., Omsland, T. K., Holvik, K., Meyer, H. E., Aamodt, G. & Norwegian Epidemiologic Osteoporosis Study Core Research, G. (2014) Do cadmium, lead, and aluminum in drinking water increase the risk of hip fractures? A NOREPOS study. *Biol Trace Elem Res* **157,** 14-23. doi:10.1007/s12011-013-9862-x.

Darveau, R. P. (2010) Periodontitis: a polymicrobial disruption of host homeostasis. *Nat Rev Microbiol* **8,** 481-490. doi:10.1038/nrmicro2337.

Dietrich, T., Walter, C., Oluwagbemigun, K., Bergmann, M., Pischon, T., Pischon, N. & Boeing, H. (2015) Smoking, Smoking Cessation, and Risk of Tooth Loss: The EPIC-Potsdam Study. *J Dent Res* **94,** 1369-1375. doi:10.1177/0022034515598961.

Ding, C., Ji, X., Chen, X., Xu, Y. & Zhong, L. (2014) TNF-alpha gene promoter polymorphisms contribute to periodontitis susceptibility: evidence from 46 studies. *J Clin Periodontol* **41,** 748-759. doi:10.1111/jcpe.12279.

Divaris, K., Monda, K. L., North, K. E., Olshan, A. F., Reynolds, L. M., Hsueh, W. C., Lange, E. M., Moss, K., Barros, S. P., Weyant, R. J., Liu, Y., Newman, A. B., Beck, J. D. & Offenbacher, S. (2013) Exploring the genetic basis of chronic periodontitis: a genome-wide association study. *Hum Mol Genet* **22,** 2312-2324. doi:10.1093/hmg/ddt065.

Dodington, D. W., Fritz, P. C., Sullivan, P. J. & Ward, W. E. (2015) Higher Intakes of Fruits and Vegetables, beta-Carotene, Vitamin C, alpha-Tocopherol, EPA, and DHA Are Positively Associated with Periodontal Healing after Nonsurgical Periodontal Therapy in Nonsmokers but Not in Smokers. *J Nutr* **145,** 2512-2519. doi:10.3945/jn.115.211524.

Eke, P. I., Wei, L., Thornton-Evans, G. O., Borrell, L. N., Borgnakke, W. S., Dye, B. & Genco, R. J. (2016) Risk Indicators for Periodontitis in US Adults: National Health and Nutrition Examination Survey (NHANES) 2009 - 2012. *J Periodontol***,** 1-18. doi:10.1902/jop.2016.160013.

El-Said KF, E.-G. A., Mahdy NH, El-Bastawy NA (2008) Chronic occupational exposure to lead and its impact on oral health. *J Egypt Public Health Assoc.* **83,** 451-466.

Ellick, R. M., Fitzgerald, S. D., Link, J. E. & Bursian, S. J. (2013) Comparison of destructive periodontal disease in blue iris mink to PCB 126-induced mandibular and maxillary squamous epithelial proliferation in natural dark mink. *Toxicol Pathol* **41,** 528-531. doi:10.1177/0192623312457270.

Feng, P., Wang, X., Casado, P. L., Kuchler, E. C., Deeley, K., Noel, J., Kimm, H., Kim, J. H., Haas, A. N., Quinelato, V., Bonato, L. L., Granjeiro, J. M., Susin, C. & Vieira, A. R. (2014) Genome wide association scan for chronic periodontitis implicates novel locus. *BMC Oral Health* **14,** 84. doi:10.1186/1472-6831-14-84.

Firdous, A. P., Kuttan, G. & Kuttan, R. (2015) Anti-inflammatory potential of carotenoid meso-zeaxanthin and its mode of action. *Pharm Biol* **53,** 961-967. doi:10.3109/13880209.2014.950673.

Gammone, M. A., Riccioni, G. & D'Orazio, N. (2015) Carotenoids: potential allies of cardiovascular health? *Food Nutr Res* **59,** 26762. doi:10.3402/fnr.v59.26762.

Grimm, F. A., Hu, D., Kania-Korwel, I., Lehmler, H. J., Ludewig, G., Hornbuckle, K. C., Duffel, M. W., Bergman, A. & Robertson, L. W. (2015) Metabolism and metabolites of polychlorinated biphenyls. *Crit Rev Toxicol* **45,** 245-272. doi:10.3109/10408444.2014.999365.

Grossi, S. G., Zambon, J. J., Ho, A. W., Koch, G., Dunford, R. G., Machtei, E. E., Norderyd, O. M. & Genco, R. J. (1994) Assessment of risk for periodontal disease. I. Risk indicators for attachment loss. *J Periodontol* **65,** 260-267. doi:10.1902/jop.1994.65.3.260.

Hajishengallis, G. (2014) Immunomicrobial pathogenesis of periodontitis: keystones, pathobionts, and host response. *Trends Immunol* **35,** 3-11. doi:10.1016/j.it.2013.09.001.

Hamers, T., Kamstra, J. H., Cenijn, P. H., Pencikova, K., Palkova, L., Simeckova, P., Vondracek, J., Andersson, P. L., Stenberg, M. & Machala, M. (2011) In vitro toxicity profiling of ultrapure non-dioxin-like polychlorinated biphenyl congeners and their relative toxic contribution to PCB mixtures in humans. *Toxicol Sci* **121,** 88-100. doi:10.1093/toxsci/kfr043.

Hennig, R., Ding, X. Z., Tong, W. G., Schneider, M. B., Standop, J., Friess, H., Buchler, M. W., Pour, P. M. & Adrian, T. E. (2002) 5-Lipoxygenase and leukotriene B(4) receptor are expressed in human pancreatic cancers but not in pancreatic ducts in normal tissue. *Am J Pathol* **161,** 421-428. doi:10.1016/S0002-9440(10)64198-3.

Hu, Q., Fu, H., Song, H., Ren, T., Li, L., Ye, L., Liu, T. & Dong, S. (2011) Low-level lead exposure attenuates the expression of three major isoforms of neural cell adhesion molecule. *Neurotoxicology* **32,** 255-260. doi:10.1016/j.neuro.2010.12.007.

Imbeault, P., Findlay, C. S., Robidoux, M. A., Haman, F., Blais, J. M., Tremblay, A., Springthorpe, S., Pal, S., Seabert, T., Krummel, E. M., Maal-Bared, R., Tetro, J. A., Pandey, S., Sattar, S. A. & Filion, L. G. (2012) Dysregulation of cytokine response in Canadian First Nations communities: is there an association with persistent organic pollutant levels? *PLoS One* **7,** e39931. doi:10.1371/journal.pone.0039931.

Jansson, H., Wahlin, A., Johansson, V., Akerman, S., Lundegren, N., Isberg, P. E. & Norderyd, O. (2014) Impact of periodontal disease experience on oral health-related quality of life. *J Periodontol* **85,** 438-445. doi:10.1902/jop.2013.130188.

Kaulmann, A. & Bohn, T. (2014) Carotenoids, inflammation, and oxidative stress--implications of cellular signaling pathways and relation to chronic disease prevention. *Nutr Res* **34,** 907-929. doi:10.1016/j.nutres.2014.07.010.

Kaye, E. K., Chen, N., Cabral, H. J., Vokonas, P. & Garcia, R. I. (2016) Metabolic Syndrome and Periodontal Disease Progression in Men. *J Dent Res* **95,** 822-828. doi:10.1177/0022034516641053.

Kim, M. J., Pelloux, V., Guyot, E., Tordjman, J., Bui, L. C., Chevallier, A., Forest, C., Benelli, C., Clement, K. & Barouki, R. (2012) Inflammatory pathway genes belong to major targets of persistent organic pollutants in adipose cells. *Environ Health Perspect* **120,** 508-514. doi:10.1289/ehp.1104282.

Kim, Y. & Lee, B. K. (2013) Increased erythrocyte lead levels correlate with decreased hemoglobin levels in the Korean general population: analysis of 2008-2010 Korean National Health and Nutrition Examination Survey data. *Int Arch Occup Environ Health* **86,** 741-748. doi:10.1007/s00420-012-0811-3.

Kumar, J., Lind, P. M., Salihovic, S., van Bavel, B., Ingelsson, E. & Lind, L. (2014) Persistent organic pollutants and inflammatory markers in a cross-sectional study of elderly Swedish people: the PIVUS cohort. *Environ Health Perspect* **122,** 977-983. doi:10.1289/ehp.1307613.

Larsson, L., Castilho, R. M. & Giannobile, W. V. (2015) Epigenetics and its role in periodontal diseases: a state-of-the-art review. *J Periodontol* **86,** 556-568. doi:10.1902/jop.2014.140559.

Larsson, L., Thorbert-Mros, S., Rymo, L. & Berglundh, T. (2012) Influence of epigenetic modifications of the interleukin-10 promoter on IL10 gene expression. *Eur J Oral Sci* **120,** 14-20. doi:10.1111/j.1600-0722.2011.00917.x.

Lavu, V., Venkatesan, V. & Rao, S. R. (2015) The epigenetic paradigm in periodontitis pathogenesis. *J Indian Soc Periodontol* **19,** 142-149. doi:10.4103/0972-124X.145784.

Levin, M., Morsey, B., Mori, C., Nambiar, P. R. & De Guise, S. (2005) Non-coplanar PCB-mediated modulation of human leukocyte phagocytosis: a new mechanism for immunotoxicity. *J Toxicol Environ Health A* **68,** 1977-1993. doi:10.1080/15287390500227126.

Linden, G. J., McClean, K. M., Woodside, J. V., Patterson, C. C., Evans, A., Young, I. S. & Kee, F. (2009) Antioxidants and periodontitis in 60-70-year-old men. *J Clin Periodontol* **36,** 843-849. doi:10.1111/j.1600-051X.2009.01468.x.

Liu, D., Perkins, J. T. & Hennig, B. (2016) EGCG prevents PCB-126-induced endothelial cell inflammation via epigenetic modifications of NF-kappaB target genes in human endothelial cells. *J Nutr Biochem* **28,** 164-170. doi:10.1016/j.jnutbio.2015.10.003.

Liu, D., Perkins, J. T., Petriello, M. C. & Hennig, B. (2015) Exposure to coplanar PCBs induces endothelial cell inflammation through epigenetic regulation of NF-kappaB subunit p65. *Toxicol Appl Pharmacol* **289,** 457-465. doi:10.1016/j.taap.2015.10.015.

Loo, W. T., Jin, L., Cheung, M. N., Wang, M. & Chow, L. W. (2010) Epigenetic change in E-cadherin and COX-2 to predict chronic periodontitis. *J Transl Med* **8,** 110. doi:10.1186/1479-5876-8-110.

Loos, B. G., Papantonopoulos, G., Jepsen, S. & Laine, M. L. (2015) What is the Contribution of Genetics to Periodontal Risk? *Dent Clin North Am* **59,** 761-780. doi:10.1016/j.cden.2015.06.005.

Lucchini, R. G. & Hashim, D. (2015) Tremor secondary to neurotoxic exposure: mercury, lead, solvents, pesticides. *Handb Clin Neurol* **131,** 241-249. doi:10.1016/B978-0-444-62627-1.00014-7.

Madupu, R., Szpakowski, S. & Nelson, K. E. (2013) Microbiome in human health and disease. *Sci Prog* **96,** 153-170.

Majkova, Z., Layne, J., Sunkara, M., Morris, A. J., Toborek, M. & Hennig, B. (2011) Omega-3 fatty acid oxidation products prevent vascular endothelial cell activation by coplanar polychlorinated biphenyls. *Toxicol Appl Pharmacol* **251,** 41-49. doi:10.1016/j.taap.2010.11.013.

Martinez-Canut P, L. A., Magan R (2005) Smoking and Periodontal disease severity. In: *J Clin Periodontol*, pp. 743-749.

Matsumoto, C., Ashida, N., Yokoyama, S., Tominari, T., Hirata, M., Ogawa, K., Sugiura, M., Yano, M., Inada, M. & Miyaura, C. (2013) The protective effects of beta-cryptoxanthin on inflammatory bone resorption in a mouse experimental model of periodontitis. *Biosci Biotechnol Biochem* **77,** 860-862. doi:10.1271/bbb.120791.

Meyle, J. & Chapple, I. (2015) Molecular aspects of the pathogenesis of periodontitis. *Periodontol 2000* **69,** 7-17. doi:10.1111/prd.12104.

Michalowicz, B. S., Diehl, S. R., Gunsolley, J. C., Sparks, B. S., Brooks, C. N., Koertge, T. E., Califano, J. V., Burmeister, J. A. & Schenkein, H. A. (2000) Evidence of a substantial genetic basis for risk of adult periodontitis. *J Periodontol* **71,** 1699-1707. doi:10.1902/jop.2000.71.11.1699.

Moon, S. S. (2013) Association of lead, mercury and cadmium with diabetes in the Korean population: the Korea National Health and Nutrition Examination Survey (KNHANES) 2009-2010. *Diabet Med* **30,** e143-148. doi:10.1111/dme.12103.

Mundy, W. R., Shafer, T. J., Tilson, H. A. & Kodavanti, P. R. (1999) Extracellular calcium is required for the polychlorinated biphenyl-induced increase of intracellular free calcium levels in cerebellar granule cell culture. *Toxicology* **136,** 27-39.

Nishigaki, M., Yamamoto, T., Ichioka, H., Honjo, K., Yamamoto, K., Oseko, F., Kita, M., Mazda, O. & Kanamura, N. (2013) beta-cryptoxanthin regulates bone resorption related-cytokine production in human periodontal ligament cells. *Arch Oral Biol* **58,** 880-886. doi:10.1016/j.archoralbio.2013.01.005.

Patel, C. J. & Manrai, A. K. (2015) Development of exposome correlation globes to map out environment-wide associations. *Pac Symp Biocomput***,** 231-242.

Petersen, P. E. & Ogawa, H. (2005) Strengthening the prevention of periodontal disease: the WHO approach. *J Periodontol* **76,** 2187-2193. doi:10.1902/jop.2005.76.12.2187.

Pihlstrom, B. L., Michalowicz, B. S. & Johnson, N. W. (2005) Periodontal diseases. *Lancet* **366,** 1809-1820. doi:10.1016/S0140-6736(05)67728-8.

Rahman, A., Brew, B. J. & Guillemin, G. J. (2011) Lead dysregulates serine/threonine protein phosphatases in human neurons. *Neurochem Res* **36,** 195-204. doi:10.1007/s11064-010-0300-6.

Reichrath, J., Lehmann, B., Carlberg, C., Varani, J. & Zouboulis, C. C. (2007) Vitamins as hormones. *Horm Metab Res* **39,** 71-84. doi:10.1055/s-2007-958715.

Reynolds, M. A. (2014) Modifiable risk factors in periodontitis: at the intersection of aging and disease. *Periodontol 2000* **64,** 7-19. doi:10.1111/prd.12047.

Rhee, S. Y., Hwang, Y. C., Woo, J. T., Sinn, D. H., Chin, S. O., Chon, S. & Kim, Y. S. (2013) Blood lead is significantly associated with metabolic syndrome in Korean adults: an analysis based on the Korea National Health and Nutrition Examination Survey (KNHANES), 2008. *Cardiovasc Diabetol* **12,** 9. doi:10.1186/1475-2840-12-9.

Rhodin, K., Divaris, K., North, K. E., Barros, S. P., Moss, K., Beck, J. D. & Offenbacher, S. (2014) Chronic periodontitis genome-wide association studies: gene-centric and gene set enrichment analyses. *J Dent Res* **93,** 882-890. doi:10.1177/0022034514544506.

Roberts, F. A. & Darveau, R. P. (2015) Microbial protection and virulence in periodontal tissue as a function of polymicrobial communities: symbiosis and dysbiosis. *Periodontol 2000* **69,** 18-27. doi:10.1111/prd.12087.

Saraiva, M. C., Taichman, R. S., Braun, T., Nriagu, J., Eklund, S. A. & Burt, B. A. (2007) Lead exposure and periodontitis in US adults. *J Periodontal Res* **42,** 45-52. doi:10.1111/j.1600-0765.2006.00913.x.

Scapoli, L., Girardi, A., Palmieri, A., Martinelli, M., Cura, F., Lauritano, D., Pezzetti, F. & Carinci, F. (2015) Interleukin-6 Gene Polymorphism Modulates the Risk of Periodontal Diseases. *J Biol Regul Homeost Agents* **29,** 111-116.

Schulz, S., Immel, U. D., Just, L., Schaller, H. G., Glaser, C. & Reichert, S. (2016) Epigenetic characteristics in inflammatory candidate genes in aggressive periodontitis. *Hum Immunol* **77,** 71-75. doi:10.1016/j.humimm.2015.10.007.

Senut, M. C., Cingolani, P., Sen, A., Kruger, A., Shaik, A., Hirsch, H., Suhr, S. T. & Ruden, D. (2012) Epigenetics of early-life lead exposure and effects on brain development. *Epigenomics* **4,** 665-674. doi:10.2217/epi.12.58.

Sipka, S., Eum, S. Y., Son, K. W., Xu, S., Gavalas, V. G., Hennig, B. & Toborek, M. (2008) ORAL ADMINISTRATION OF PCBs INDUCES PROINFLAMMATORY AND PROMETASTATIC RESPONSES. *Environ Toxicol Pharmacol* **25,** 251-259. doi:10.1016/j.etap.2007.10.020.

Song, H., Zheng, G., Liu, Y., Shen, X. F., Zhao, Z. H., Aschner, M., Luo, W. J. & Chen, J. Y. (2016) Cellular uptake of lead in the blood-cerebrospinal fluid barrier: Novel roles of Connexin 43 hemichannel and its down-regulations via Erk phosphorylation. *Toxicol Appl Pharmacol* **297,** 1-11. doi:10.1016/j.taap.2016.02.021.

Spinas, E., Saggini, A., Kritas, S. K., Cerulli, G., Caraffa, A., Antinolfi, P., Pantalone, A., Frydas, A., Tei, M., Speziali, A., Saggini, R., Pandolfi, F. & Conti, P. (2015) Can vitamin a mediate immunity and inflammation? *J Biol Regul Homeost Agents* **29,** 1-6.

Stabholz, A., Soskolne, W. A. & Shapira, L. (2010) Genetic and environmental risk factors for chronic periodontitis and aggressive periodontitis. *Periodontol 2000* **53,** 138-153. doi:10.1111/j.1600-0757.2010.00340.x.

Strobl, C., Malley, J. & Tutz, G. (2009) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* **14,** 323-348. doi:10.1037/a0016973.

Tareke, E., Rydberg, P., Karlsson, P., Eriksson, S. & Tornqvist, M. (2002) Analysis of acrylamide, a carcinogen formed in heated foodstuffs. *J Agric Food Chem* **50,** 4998-5006.

Terrizzi, A. R., Fernandez-Solari, J., Lee, C. M., Bozzini, C., Mandalunis, P. M., Elverdin, J. C., Conti, M. I. & Martinez, M. P. (2013) Alveolar bone loss associated to periodontal disease in lead intoxicated rats under environmental hypoxia. *Arch Oral Biol* **58,** 1407-1414. doi:10.1016/j.archoralbio.2013.06.010.

Terrizzi, A. R., Fernandez-Solari, J., Lee, C. M., Martinez, M. P. & Conti, M. I. (2014) Lead intoxication under environmental hypoxia impairs oral health. *J Toxicol Environ Health A* **77,** 1304-1310. doi:10.1080/15287394.2014.938209.

Thurnham, D. I., Northrop-Clewes, C. A. & Knowles, J. (2015) The use of adjustment factors to address the impact of inflammation on vitamin A and iron status in humans. *J Nutr* **145,** 1137S-1143S. doi:10.3945/jn.114.194712.

Tilson, H. A. (1998) Developmental neurotoxicology of endocrine disruptors and pesticides: identification of information gaps and research needs. *Environ Health Perspect* **106 Suppl 3,** 807-811.

Tomar, S. L. & Asma, S. (2000) Smoking-attributable periodontitis in the United States: findings from NHANES III. National Health and Nutrition Examination Survey. *J Periodontol* **71,** 743-751. doi:10.1902/jop.2000.71.5.743.

Vaiserman, A. (2014) Early-life Exposure to Endocrine Disrupting Chemicals and Later-life Health Outcomes: An Epigenetic Bridge? *Aging Dis* **5,** 419-429. doi:10.14336/AD.2014.0500419.

Van Dyke, T. E. & Sheilesh, D. (2005) Risk factors for periodontitis. *J Int Acad Periodontol* **7,** 3-7.

Vesper, H. W., Bernert, J. T., Ospina, M., Meyers, T., Ingham, L., Smith, A. & Myers, G. L. (2007) Assessment of the relation between biomarkers for smoking and biomarkers for acrylamide exposure in humans. *Cancer Epidemiol Biomarkers Prev* **16,** 2471-2478. doi:10.1158/1055-9965.EPI-06-1058.

Wang, L., Wang, J., Jin, Y., Gao, H. & Lin, X. (2014) Oral administration of all-trans retinoic acid suppresses experimental periodontitis by modulating the Th17/Treg imbalance. *J Periodontol* **85,** 740-750. doi:10.1902/jop.2013.130132.

Won, Y. S., Kim, J. H., Kim, Y. S. & Bae, K. H. (2013) Association of internal exposure of cadmium and lead with periodontal disease: a study of the Fourth Korean National Health and Nutrition Examination Survey. *J Clin Periodontol* **40,** 118-124. doi:10.1111/jcpe.12033.

Wu, X., Offenbacher, S., Lomicronpez, N. J., Chen, D., Wang, H. Y., Rogus, J., Zhou, J., Beck, J., Jiang, S., Bao, X., Wilkins, L., Doucette-Stamm, L. & Kornman, K. (2015) Association of interleukin-1 gene variations with moderate to severe chronic periodontitis in multiple ethnicities. *J Periodontal Res* **50,** 52-61. doi:10.1111/jre.12181.

Xu, L., Zhu, J., Yin, W. & Ding, X. (2015) Astaxanthin improves cognitive deficits from oxidative stress, nitric oxide synthase and inflammation through upregulation of PI3K/Akt in diabetes rat. *Int J Clin Exp Pathol* **8,** 6083-6094.

Zhang, N., Baker, H. W., Tufts, M., Raymond, R. E., Salihu, H. & Elliott, M. R. (2013) Early childhood lead exposure and academic achievement: evidence from Detroit public schools, 2008-2010. *Am J Public Health* **103,** e72-77. doi:10.2105/AJPH.2012.301164.

Zhang, S., Barros, S. P., Niculescu, M. D., Moretti, A. J., Preisser, J. S. & Offenbacher, S. (2010) Alteration of PTGS2 promoter methylation in chronic periodontitis. *J Dent Res* **89,** 133-137. doi:10.1177/0022034509356512.

Table 2.1: Demographic information by periodontal status.

	N	No Periodontitis	Periodontitis	p-value	Weighted N	Weighted Prevalence of Periodontitis	Weighted p-value
N	8884	7915 (89.1%)	969 (10.9%)	.	137140007	8.1%	.
Male	4297 (48.4%)	3706 (86.2%)	591 (13.8%)	<.0001	68191901	10.1%	<.0001
Female	4587 (51.6%)	4209 (91.8%)	378 (8.2%)	.	68948106	6.2%	.
Mexican American	2233 (25.1%)	1913 (85.7%)	320 (14.3%)	<.0001	11566835	12.3%	<.0001
Other Hispanic	417 (4.7%)	355 (85.1%)	62 (14.9%)	.	8127478	14.6%	.
Non-Hispanic White	4305 (48.5%)	4030 (93.6%)	275 (6.4%)	.	97058220	5.8%	.
NonHispanic Black	1603 (18.0%)	1328 (82.8%)	275 (17.2%)	.	13910439	16.0%	.
Other race	326 (3.7%)	289 (88.7%)	37 (11.3%)	.	6477034	10.3%	.
Age 18~30	2497 (28.1%)	2390 (95.7%)	107 (4.3%)	<.0001	36006214	3.1%	<.0001
Age 31~49	3522 (39.6%)	3068 (87.1%)	454 (12.9%)	.	63781453	9.3%	.
Age 50~64	1667 (18.8%)	1435 (86.1%)	232 (13.9%)	.	26103632	10.9%	.
Age 65+	1198 (13.5%)	1022 (85.3%)	176 (14.7%)	.	11248708	11.6%	.
Age (Mean, StdErr)	43.12 (0.18)	42.47 (0.19)	48.47 (0.49)	<.0001	41.55 (0.26)	46.51 (0.53)	<.0001
Non Smoker	4967 (55.9%)	4538 (91.4%)	429 (8.6%)	<.0001	73898456	5.8%	<.0001
Current Smoker	1920 (21.6%)	1616 (84.2%)	304 (15.8%)	.	32664766	13.4%	.
Former Smoker	1997 (22.5%)	1761 (88.2%)	236 (11.8%)	.	30576785	8.0%	.
Socio-Eco Status (Mean, StdErr)	2.75 (0.02)	2.82 (0.02)	2.19 (0.05)	<.0001	3.12 (0.05)	2.50 (0.08)	<.0001
Totalteeth (Mean, StdErr)	26.19 (0.04)	26.29 (0.04)	25.38 (0.13)	<.0001	26.26 (0.06)	25.12 (0.17)	<.0001

Table 2.2: Demographic information by smoking status.

	N	Non Smoker	Current Smoker	Former Smoker	p-value	Weighted N	Weighted Non Smoker	Weighted Current Smoker	Weighted Former Smoker	Weighted p-value
N	8884	4967 (55.9%)	1920 (21.6%)	1997 (22.5%)	.	137140007	53.9%	23.8%	22.3%	.
Male	4297 (48.4%)	2020 (47.0%)	1119 (26.0%)	1158 (26.9%)	<.0001	68191901	48.1%	26.7%	25.2%	<.0001
Female	4587 (51.6%)	2947 (64.2%)	801 (17.5%)	839 (18.3%)	.	68948106	59.6%	20.9%	19.4%	.
Mexican American	2233 (25.1%)	1335 (59.8%)	413 (18.5%)	485 (21.7%)	<.0001	11566835	58.3%	22.6%	19.2%	<.0001
Other Hispanic	417 (4.7%)	243 (58.3%)	91 (21.8%)	83 (19.9%)	.	8127478	55.6%	24.6%	19.7%	.
Non-Hispanic White	4305 (48.5%)	2227 (51.7%)	935 (21.7%)	1143 (26.6%)	.	97058220	51.7%	23.6%	24.7%	.
Non-Hispanic Black	1603 (18.0%)	963 (60.1%)	408 (25.5%)	232 (14.5%)	.	13910439	61.9%	25.8%	12.3%	.
Other race	326 (3.7%)	199 (61.0%)	73 (22.4%)	54 (16.6%)	.	6477034	59.2%	24.7%	16.1%	.
Age 18~30	2497 (28.1%)	1542 (61.8%)	652 (26.1%)	303 (12.1%)	<.0001	36006214	57.3%	31.6%	11.1%	<.0001
Age 31~49	3522 (39.6%)	1932 (54.9%)	923 (26.2%)	667 (18.9%)	.	63781453	53.7%	25.6%	20.7%	.
Age 50~64	1667 (18.8%)	827 (49.6%)	281 (16.9%)	559 (33.5%)	.	26103632	47.7%	17.2%	35.1%	.
Age 65+	1198 (13.5%)	666 (55.6%)	64 (5.3%)	468 (39.1%)	.	11248708	58.1%	4.5%	37.4%	.
Age (Mean, StdErr)	43.12 (0.18)	42.20 (0.24)	37.95 (0.29)	50.38 (0.38)	<.0001	41.55 (0.26)	41.09 (0.35)	36.82 (0.33)	47.71 (0.41)	<.0001
Socio-Eco Status (Mean, StdErr)	2.75 (0.02)	2.81 (0.02)	2.26 (0.04)	3.09 (0.04)	<.0001	3.12 (0.05)	3.21 (0.06)	2.56 (0.07)	3.48 (0.06)	<.0001
Totalteeth (Mean, StdErr)	26.19 (0.04)	26.55 (0.05)	26.05 (0.08)	25.44 (0.08)	<.0001	26.26 (0.06)	26.63 (0.06)	25.98 (0.12)	25.67 (0.10)	<.0001

Table 2.3 Directionality of gender, race and age to prevalence of periodontal disease, n=8884.

	Name	Percentage of Periodontitis	Name	Percentage of Periodontitis	
1	Male	13.8%	Female	8.2%	<.0001
2	Mexican American	14.3%	Other Hispanic	14.9%	.3817
3	Mexican American	14.3%	NonHispanic White	6.4%	<.0001
4	Mexican American	14.3%	NonHispanic Black	17.2%	.0086
5	Mexican American	14.3%	Other race	11.3%	.0734
6	Other Hispanic	14.9%	NonHispanic White	6.4%	<.0001
7	Other Hispanic	14.9%	NonHispanic Black	17.2%	.1322
8	Other Hispanic	14.9%	Other race	11.3%	.0807
9	NonHispanic White	6.4%	NonHispanic Black	17.2%	<.0001
10	NonHispanic White	6.4%	Other race	11.3%	.0003
11	NonHispanic Black	17.2%	Other race	11.3%	.0047
12	Age 18~30	4.3%	Age 31~49	12.9%	<.0001
13	Age 18~30	4.3%	Age 50~64	13.9%	<.0001
14	Age 18~30	4.3%	Age 65+	14.7%	<.0001
15	Age 31~49	12.9%	Age 50~64	13.9%	.1646
16	Age 31~49	12.9%	Age 65+	14.7%	.0625
17	Age 50~64	13.9%	Age 65+	14.7%	.2979

Table 2.4: Environmental variables and parameter estimates with significant survey weighted logistic regressions. This table presents subset of environmental factors as risk factor for periodontal disease. Environmental variables included are those with p-value <0.01. OR are listed in descending order.

Environmental Factor	Class	Odds Ratio	95 % CI	SEM	p
PCB206 (ng/g)	pcb	2.65	(1.88, 3.75)	0.177	<0.0001
PCB172 (ng/g)	pcb	2.18	(1.61, 2.97)	0.156	<0.0001
PCB157 (ng/g)	pcb	2.05	(1.46, 2.87)	0.173	0.0002
PCB178 (ng/g)	pcb	2.00	(1.50, 2.66)	0.145	<0.0001
PCB177 (ng/g)	pcb	1.99	(1.53, 2.60)	0.135	<0.0001
PCB199 (ng/g)	pcb	1.96	(1.29, 2.97)	0.213	0.0046
PCB183 (ng/g)	pcb	1.84	(1.47, 2.29)	0.113	<0.0001
PCB194 (ng/g)	pcb	1.82	(1.23, 2.69)	0.200	0.0068
PCB196 & 203 (ng/g)	pcb	1.70	(1.18, 2.43)	0.184	0.0087
PCB170 (ng/g)	pcb	1.69	(1.28, 2.23)	0.141	0.0007
PCB167 (ng/g)	pcb	1.69	(1.26, 2.27)	0.150	0.0012
Lead (µg/dL)	heavy metal	1.66	(1.47, 1.87)	0.062	<0.0001
2-fluorene (ng/L)	hydrocarbon	1.64	(1.38, 1.94)	0.086	<0.0001
3-fluorene (ng/L)	hydrocarbon	1.63	(1.35, 1.96)	0.095	<0.0001
Benzene (ng/mL)	volatile compound	1.63	(1.27, 2.10)	0.128	0.0006
Cotinine (ng/mL)	alkaloid	1.56	(1.39, 1.75)	0.058	<0.0001
PCB153 (ng/g)	pcb	1.56	(1.18, 2.06)	0.141	0.0033
Cadmium (µg/L)	heavy metal	1.54	(1.41, 1.68)	0.046	<0.0001
PCB187 (ng/g)	pcb	1.53	(1.19, 1.97)	0.129	0.0023
PCB156 (ng/g)	pcb	1.52	(1.13, 2.05)	0.152	0.0093
Toluene (ng/mL)	volatile compound	1.51	(1.20, 1.88)	0.114	0.0010
PCB146 (ng/g)	pcb	1.51	(1.18, 1.94)	0.126	0.0023
PCB105 (ng/g)	pcb	1.49	(1.22, 1.82)	0.101	0.0003
Cadmium, urine (ng/mL)	heavy metal	1.47	(1.16, 1.85)	0.120	0.0029
1-pyrene (ng/L)	hydrocarbon	1.46	(1.19, 1.79)	0.105	0.0015
1-napthol (ng/L)	hydrocarbon	1.43	(1.22, 1.69)	0.083	0.0003
PCB66 (ng/g)	pcb	1.43	(1.18, 1.72)	0.096	0.0007
2-napthol (ng/L)	hydrocarbon	1.42	(1.18, 1.70)	0.094	0.0013
2,3,7,8-tcdd (fg/g)	dioxins	1.41	(1.19, 1.68)	0.089	0.0004
2-phenanthrene (ng/L)	hydrocarbon	1.41	(1.14, 1.74)	0.109	0.0049
1-phenanthrene (ng/L)	hydrocarbon	1.38	(1.17, 1.63)	0.085	0.0011
3-phenanthrene (ng/L)	hydrocarbon	1.36	(1.15, 1.61)	0.085	0.0015
Styrene (ng/mL)	volatile compound	1.36	(1.10, 1.69)	0.111	0.0083
Antimony, urine (ng/mL)	heavy metal	1.28	(1.12, 1.45)	0.064	0.0006
Retinyl stearate (µg/dL)	nutrient	1.19	(1.08, 1.32)	0.052	0.0016
α-tocopherol (µg/dL)	nutrient	1.16	(1.05, 1.27)	0.050	0.0061

Folate, RBC (ng/mL RBC)	nutrient	0.85	(0.76, 0.94)	0.055	0.0040
Vitamin D (ng/mL)	nutrient	0.83	(0.73, 0.93)	0.060	0.0047
trans-β-carotene (μg/dL)	nutrient	0.81	(0.70, 0.93)	0.073	0.0083
Folate, serum (ng/mL)	nutrient	0.80	(0.71, 0.90)	0.062	0.0010
β-cryptoxanthin (μg/dL)	nutrient	0.80	(0.71, 0.91)	0.062	0.0021
α-Carotene (μg/dL)	nutrient	0.80	(0.68, 0.93)	0.081	0.0097

Table 2.5: Environmental variables and the parameter estimates from survey-weighted logistic regressions stratified by smoking groups

		Current Smokers			Former Smokers			Non-Smokers		
		Odds	**95% CI**	**p**	**Odds**	**95% CI**	**p**	**Odds**	**95% CI**	**p**
Lead	Heavy metal	1.54	(1.28, 1.87)	<0.001	1.57	(1.27, 1.94)	<0.001	1.39	(1.18, 1.65)	<0.001
PCB105 (ng/g)	Pcb	1.68	(1.28, 2.2)	0.001	1.89	(1.39, 2.57)	<0.001	1.41	(1.11, 1.78)	0.007
PCB157 (ng/g)	Pcb	2.3	(1.31, 4.04)	0.006	4.66	(2.14, 10.13)	<0.001	1.77	(1.23, 2.55)	0.004
PCB172 (ng/g)	Pcb	2.59	(1.6, 4.21)	0.001	4.42	(2.12, 9.23)	<0.001	1.69	(1.19, 2.4)	0.006
PCB177 (ng/g)	Pcb	2.18	(1.42, 3.33)	0.001	3.12	(1.69, 5.75)	0.001	1.75	(1.24, 2.46)	0.003
PCB178 (ng/g)	Pcb	2.65	(1.65, 4.25)	<0.001	2.78	(1.59, 4.86)	0.001	1.58	(1.15, 2.16)	0.008
PCB206 (ng/g)	Pcb	3.96	(1.96, 8.01)	0.001	5.29	(1.68, 16.65)	0.009	1.96	(1.3, 2.94)	0.004
PCB183 (ng/g)	Pcb	2.23	(1.54, 3.22)	<0.001	2.03	(1.23, 3.33)	0.008	1.53	(1.12, 2.09)	0.011
Mono-n-methyl phthalate	Phthalate	1.47	(1.13, 1.92)	0.01	0.71	(0.44, 1.15)	0.178	0.99	(0.75, 1.31)	0.948
Cadmium (µg/L)	Heavy metal	1.32	(1.09, 1.58)	0.006	1.32	(0.96, 1.83)	0.097	1.26	(1.03, 1.55)	0.032
1,2,3,7,8-pncdd (fg/g)	Dioxins	1.66	(1.16, 2.37)	0.008	1.09	(0.71, 1.67)	0.711	1.01	(0.76, 1.34)	0.96
2,3,7,8-tcdd (fg/g)	Dioxins	1.81	(1.21, 2.7)	0.006	1.63	(1.13, 2.35)	0.013	1.22	(0.99, 1.49)	0.07
PCB146 (ng/g)	Pcb	1.71	(1.23, 2.4)	0.003	1.87	(1.13, 3.09)	0.019	1.23	(0.84, 1.79)	0.3
PCB167 (ng/g)	Pcb	1.77	(1.23, 2.57)	0.005	2.93	(1.29, 6.7)	0.015	1.61	(1.11, 2.34)	0.017
PCB170 (ng/g)	Pcb	1.98	(1.35, 2.92)	0.001	1.36	(0.67, 2.76)	0.398	1.2	(0.84, 1.73)	0.317
PCB187 (ng/g)	Pcb	1.75	(1.19, 2.56)	0.008	1.47	(0.76, 2.82)	0.258	1.15	(0.92, 1.45)	0.233
Retinyl palmitate (µg/dL)	Nutrient	1.35	(1.09, 1.67)	0.009	0.98	(0.77, 1.24)	0.849	1.03	(0.88, 1.22)	0.701
Retinyl stearate (µg/dL)	Nutrient	1.32	(1.09, 1.59)	0.006	1.28	(1.05, 1.57)	0.021	1.07	(0.92, 1.25)	0.371
PCB66 (ng/g)	Pcb	1.63	(1.2, 2.21)	0.004	1.83	(1.4, 2.38)	<0.001	1.24	(0.95, 1.62)	0.119
Vitamin D (ng/mL)	Nutrient	1.15	(0.91, 1.45)	0.266	0.61	(0.5, 0.74)	<0.001	0.76	(0.67, 0.87)	0.001
PCB28 (ng/g)	Pcb	1.06	(0.59, 1.89)	0.853	1.9	(1.25, 2.9)	0.007	1.7	(1.12, 2.58)	0.021
Acrylamide (pmoL/G Hb)	Acrylamide	1.46	(1.06, 2.03)	0.055	0.3	(0.2, 0.45)	0.001	0.77	(0.54, 1.1)	0.195
Mono-n-octyl phthalate	Phthalate	1.23	(0.91, 1.68)	0.191	1.41	(1.1, 1.8)	0.01	1.06	(0.73, 1.54)	0.773
cis-b-carotene(µg/dL)	Nutrient	0.95	(0.81, 1.13)	0.593	1	(0.73, 1.38)	0.979	0.78	(0.67, 0.92)	0.005
Antimony, urine (ng/mL)	Heavy metal	0.87	(0.65, 1.16)	0.342	1.45	(1.05, 1.98)	0.028	1.51	(1.27, 1.8)	<0.001
Diethylphosphate (µg/L)	Organophosphates	0.76	(0.62, 0.94)	0.015	0.8	(0.57, 1.11)	0.194	1.57	(1.24, 1.99)	0.001

1,2,3,4,6,7,8,9-ocdd (fg/g)	Dioxins	1.22	(0.89, 1.67)	0.233	1.12	(0.55, 2.27)	0.755	1.46	(1.12, 1.91)	0.008

The p-values are calculated based on the survey weighted logistic regression with dichotomous periodontitis status as the outcome adjusting for age, gender, ethnicity, socioeconomic status and number of teeth. The p-value of 0.01 is determined based on the overall FDR of 5%. A significance of 0.01 corresponds to a FDR of 8%, 10%, and 13% for the current, former, and non-smokers cohort, respectively. The highlighted cells are p-values that are smaller than 0.01.

FIGURE 2 LEGENDS

Figure 2.1-2.3: Provides a Classification and Regression Tree (CART) analysis of the characteristics of discriminatory environmental variables in smokers, former smokers, and non-smokers. The CART analysis emphasizes identification of critical variables that sequentially discriminate subsets of patients with periodontitis in the 3 groups. The smoking population was represented by 1920 subject with 304 periodontitis cases; former smokers 1997 subjects with 228 cases; and non-smokers 4967 with 731 cases.

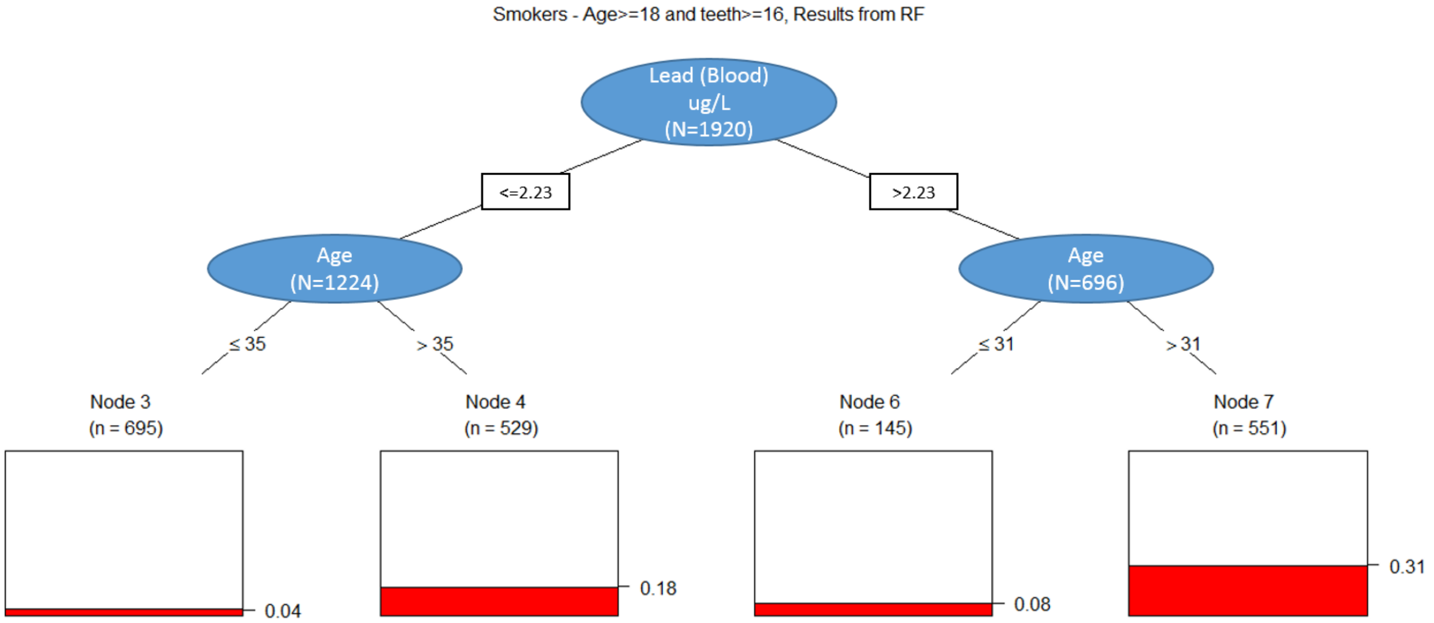


Figure 1. Cart Analysis on smoker population

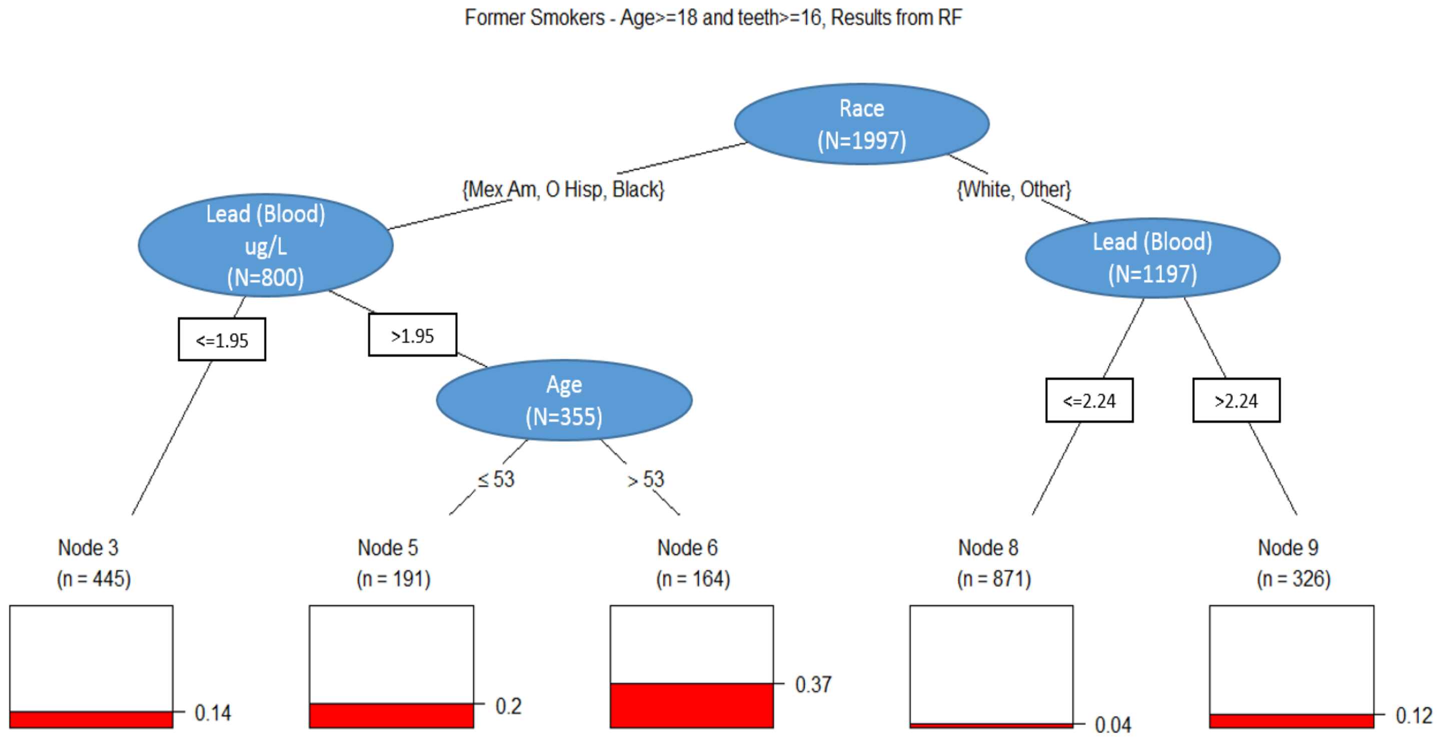


Figure 2. Cart Analysis on former smoker population

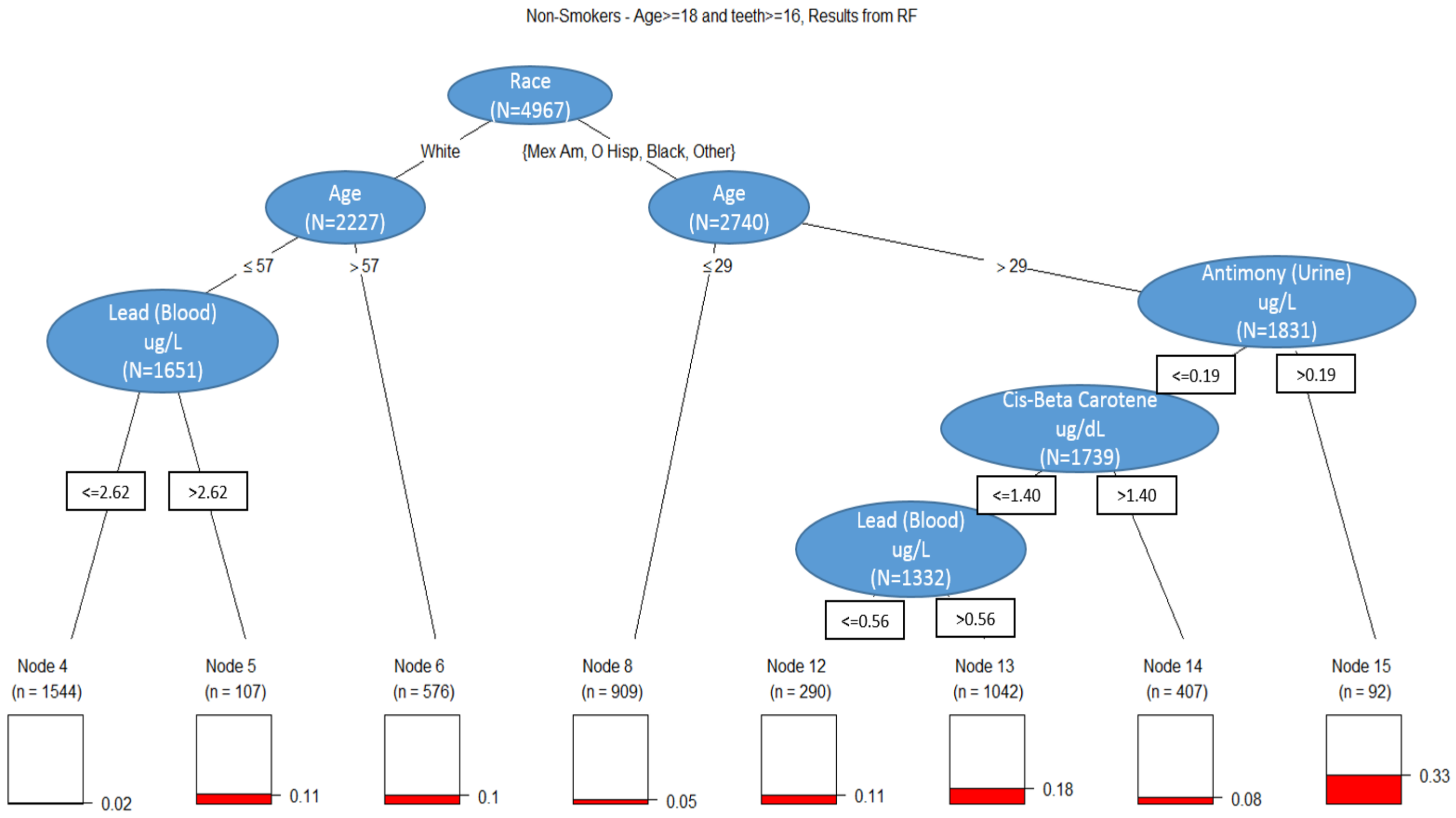


Figure 3. Cart Analysis on non-smoker population

Chapter 3 rFSA: An R Package for Finding Best Subsets and Interactions

This paper presents an **R** package that applies an algorithm intended to improve statistical models. The algorithm searches the data space for models of a specified form that are statistically optimal. Many replications of this algorithm will produce a set of 'feasible solutions', which the researcher can investigate. The algorithm can help improve existing models used in bioinformatics, health care, or other fields which have yet to explore quadratic terms, interactions, or a higher order of predictors because of the size of their datasets. The package, rFSA, is flexible for many different model forms and criteria functions. Currently, linear models and generalized linear models are supported.

3.1 Introduction

In recent years, novel statistical analysis modeling techniques and algorithms have become more computationally intensive due to advances in data mining and genetic sequencing, among other reasons. As data sizes continue to grow, the future will only be more computationally demanding. While computers have become faster, the complexity and size of the datasets have grown faster than these new computers can handle in a timely manner [89]. Usually, some level of data reduction is performed (e.g., via Principle Component Analysis, Partial Least Squares, Factor Analysis, LASSO, etc.) in order to fit statistical models and estimate coefficients. This leaves researchers and statisticians alike attempting to interpret coefficients from highly complicated statistical methods, exposing a limitation of these methods. Our method provides an alternative to data reduction by using an algorithm to improve variable selection in standard statistical models, which yields interpretable coefficient estimates and predictions.

Our algorithm, addresses the problems described above by searching a reduced space in an efficient way to produce "Feasible Solutions". Feasible Solutions [90] are optimal in the sense that no single swap of an explanatory variable in the model for a variable outside the model can improve a specified criterion function. [91] introduced this idea of a sequential replacement algorithm. According to [91], this "cheap method" yielded improved results compared to forward and backward selection, converged rapidly, and had many variations in how the algorithm could be constructed. One consequence that Miller noted was that this replacement algorithm could give too many solutions if repeated. [90], however, used this type of exchange algorithm to find minimum volume ellipsoid estimators in multivariate data and robust regression estimators. These solutions, according to [90], could provide the optimal solution with arbitrarily high probability with a sufficient number of random starts. [90] also notes this method's good performance compared to exhaustive search for the standard datasets that were tested.

This Feasible Solution Algorithm (FSA) has been implemented into a **R** package and is now accessible via GitHub.

3.1.1 Feasible Solution Algorithm.

Data analysts are often faced with the problem of identifying a subset of k explanatory variables from p variables $\mathbf{X}^p$, including interactions and quadratic terms. Consider fixing $p^+ \geq 0$ explanatory variables in a preliminary model. Denote these variables $\mathbf{X}^{p+}$. Let $m(\mathbf{Y}, \mathbf{X}^{p+})$ be an objective function that can be a measure of model quality i.e., R^2, AIC, BIC, etc. We wish to find the k additional variables denoted $\mathbf{X}^k$ to add to the model that optimizes the objective function $m(\mathbf{Y}, \mathbf{X}^{p+}, \mathbf{X}^k)$.

The Feasible Solution Algorithm (FSA) addresses this problem in the following way:

1. Choose $\mathbf{X}^k$ randomly and compute the objective function m.

2. Consider exchanging one of the k selected variables from the current model

3. Make the single exchange that improves the objective function m the most.

4. Keep making exchanges until the objective function does not improve. These variables $\mathbf{X}^{p+}, \mathbf{X}^{k^*}$ are called a feasible solution.

5. Return to (1) to find another feasible solution.

In another instance of the FSA, we include the j^{th} order interaction and lower order terms we are considering in step 1. We then continue on to step 2, only this time when we make an exchange it changes the j^{th} order interaction and the lower interactions and main effects as well. We could then optimize based on a model criterion or on an interaction terms p-value.

A single iteration of FSA yields a feasible solution in the sense that it may globally optimize $m(Y, X)$. Of course, the algorithm may converge somewhere other than the global optimum. Using the algorithm multiple times identifies multiple feasible solutions, the best of which may be the global optimum. However, the group of feasible solutions may provide useful insights into the data because each feasible solution will be unique.

[91] outlines the FSA described above in the following way. Suppose we have 26 predictor variables labeled A through Z. Imagine you wish to find the best subset with four predictor variables. First start randomly with four predictors. Suppose these are $ABCD$. Consider changing one of A, B, C, or D with one of the other 22 remaining variables. Make the change that improves the objective function the most. Suppose we swap C for X. Now we have $ABXD$. Next consider changing one of A, B, X, or D (Considering X here is redundant and not necessary). This process is repeated until no further improvements, to the objective function, can be made.

This method, coupled with repeating it for different random starts, can give different solutions which could be interesting from a clinical or scientific viewpoint. These unique feasible solutions are optimal for the criterion function that was chosen by the user in that no one exchange of any one variable can improve the criterion function.

3.1.2 Other Algorithms for Subset Regression

Many algorithms and methods exist for subset selection. Forward Selection, Backward Selection, and Stepwise Selection are common automatic variable selection techniques. Ridge Regression is a common penalized regression technique that is used for subset selection with many variables. Exhaustive Search checks all possible combinations for the possible model structures. When there are many explanatory variables to consider the exhaustive search method can be very time and resource intensive on even the fastest and most powerful computers [89]. These are some of the most common algorithms and methods that exist to find the best subset of predictors that adequately explain the response variable. These are currently available in the form of **R** packages for linear and generalized linear models in **leaps** [92], and **glmnet** [93].

3.1.3 Other Algorithms for Finding Interactions

Potential interactions can be explored via Random Forest [94] and Boosting [95] methods. Branches of the Tree are explored, and researchers can usually find places where splits in variables classify the response variable differently further into the tree. Bayesian methods also exist for exploring potential interactions in large genetic datasets [96].

Also, interactions can be searched for on an individual level by the statistician or data scientist. This is usually based on previous knowledge of the data, or expertise of the primary investigator. These processes can be tedious and can lead to interactions being ignored or missed due to the large number of interactions to check. Exhaustively searching for the optimal model that includes interactions is not always computationally possible or feasible.

Many analyses have been published which have not explored interactions. By exploring interactions, researchers will gain predictive power in their statistical models as well as uncover new and interesting insights. Time commitment for finding, and difficulty interpreting interactions limits the types of models that statisticians are able to consider. **rFSA** addresses this limitation through its ease of use and flexibility in an **R** package.

3.2 rFSA

rFSA implements the FSA algorithm described in 1.1 above for use in subset selection, and interaction finding. **rFSA** has two main functions for fitting models. **lmFSA** and **glmFSA** are the FSA analogs to the base **R** functions **lm** and **glm**. More specifically, **lmFSA** and **glmFSA**, use the base **lm** and **glm** functions to fit the necessary models for the algorithm to converge. These are made based on user inputs that describe the number of terms to consider, whether interactions or quadratic terms should be included, and the criterion function to either minimize or maximize.

3.2.1 How it is Built

In its essence, **rFSA** 's two main functions, **lmFSA** and **glmFSA** exist as tools to make the possible formula combinations to consider the model form the user has specified and then execute them via the built in **lm** and **glm** functions in R. These combinations are computed based on random starting positions described in 1.1 above and the possible swaps to consider to move. These combinations are then run on either one core or multiple cores (the number is specified by the user) via the mclapply function from the **parallel** package in R.

mclapply only supports more than one core in Unix environments. Therefore, windows users must always leave **cores=1** for **lmFSA** or **glmFSA**, while Unix users can specify more cores if their computers have more than one. If a windows user specifies more than one core, the code will change to respecify this parameter as one. We recommend using no more than one less than the maximum number of cores for your desktop computer while running either **lmFSA** or **glmFSA**. The user can run **parallel::detectCores()** to determine the number of cores on their personal machine.

lmFSA and **glmFSA** are written as **S3** objects. Returned results are of class definition, **FSA**, and can be used along side standard **S3** functions **print**, **summary**, **predict**, **fitted**, and **plot**. The **print** command on a **FSA** class will show a table of the original model specified by the user and the feasible solutions that the algorithm found from **m** random starts along with their criterion values and times they were a feasible solution. Note: Original fit is not a feasible solution, so **NA** will be listed for original fit under the ”Times.FS” column. The **summary** command on a **FSA** class will show the summary output of the original fit and feasible solutions in a list, as if you were to compute the **summary** of a **lm** or **glm** object. The **predict** and **fitted** work in a similar fashion to the **summary** function for **FSA** classes and return a list of either predicted or fitted values from the original fit and feasible solutions found. Finally, the **plot** function returns diagnostic plots of the original fit and the feasible solutions in a compact manner.

3.2.2 How it Works

rFSA has two main functions: **lmFSA** and **glmlm**. These two functions have very similar parameters. The only argument that is unique to **glmFSA** is the ”family” parameter. The ”family” parameter is used to name the family function used in the **glm** function in base R. The data that is used in **lmFSA** or **glmFSA** should contain one response variable and the rest of the predictor variables of interest. Variables known to not be of interest should be removed from the dataset prior to running either function. Data fed into **lmFSA** or **glmFSA** should be made sure that categorical variables are either listed in quotes, or factors, and that categorical variables have at least two different levels. The ”formula” parameter is where the user specifies an original fit. Here, the user specifies a response variable that will be used throughout for all model fits. If the user is unsure about what to use for the formula, they should make sure to specifying the correct response (eg. **Y**), and can simply write a simple model with just an intercept (eg. **Y 1**).

Arbitrary column names via the **colnames** function should be added before hand so results are interpretable to the user. If the statistician using **rFSA** is interested in doing only subset selection, without interactions being considered, then the parameter **interactions** must be set to FALSE. After the parameters of the function have been set and the function has ran, the random starts will "converge" on the **numrs** number of solutions. These solutions, may repeat, and will be based on where **rFSA** randomly started for each **numrs** and what criterion function was used. A list of the criteria functions that are currently supported are listed below. Individuals own criteria functions can be used as long as they follow a similar format to the hard coded criteria functions included in the **rFSA** package.

A list of five values from a successful run of **lmFSA** or **glmFSA** is attached to every **FSA** object for the users convenience. First, the original model fit ($**originalfit**) from **lm** or **glm** is given. Second, the **lmFSA** or **glmFSA** function parameters ($**call**) are given. Third, a table of each random start and the place it converged ($**solutions**). Forth, a table of a summary of the solutions and how many times each were repeated ($**table**). Fifth and finally, a printout summary of the comparison of the overall number of models that were checked by **lmFSA** or **glmFSA** and how many would have been checked by exhaustive search is returned ($**efficiency**).

Arguments:

formula	a symbolic description of the original model to be fitted.
data	a data frame containing the variables in the model.
quad	to include quadratic terms or not.
numrs	number of random starts to perform.
cores	number of cores to use while running. Note: Windows can only use 1 core.
interactions	T or F for whether to include interactions in model. Defaults to TRUE.
criterion	which criterion function to either maximize or minimize.
minmax	whether to minimize or maximize the criterion function.
family	family argument passed to **glm**.
fixvar	a variable to fix in the model. Usually a covariate(s) that should always be included.
m	order of terms to include. If interactions is TRUE then m is the order of the interaction.
...	arguments to be passed to the **lm** or **glm** function.

Criteria Functions:

Both **lmFSA** and **glmFSA** can use **apress** (Allen's Press Statistic), **int.p.val** (Interaction p-value), **AIC**, or **BIC**. **lmFSA** can also use **r.squared**, and **adj.r.squared**. **glmFSA** can use a function called **bdist** which is useful when you have a binary response and wish to explore two way interactions with only continuous explanatory variables. Specifically, this function computes the Bhattacharyya Distance [97]. When **bdist** is chosen, the Bhattacharyya Distance is computed, which is faster than the other criteria functions. For this reason, the Bhattacharyya Distance can be useful when understanding the relationships in large genetic datasets with continuous explanatory variables and a binary response.

Returned Values

$**originalfit lm** or **glm** object from the users specification from the formula parameter.

$**call** list of the **lmFSA** or **glmFSA** function parameters used.

$**solutions** data frame of fixed terms, start position, feasible solution, criterion function value (e.g., p-value of interaction), swaps to solution.

$**table** data frame of the unique feasible solutions and how many times they occurred out of the number of random starts chosen.

$**efficiency** number of models check if you had done exhaustive search versus the number checked by **lmFSA** or **glmFSA**.

Using **rFSA** is straight forward. To run a basic linear regression example see the code below. A simple logistic regression example follows as well.

```
#Linear Regression Example
#use mtcars package see help(mtcars)
data(mtcars)
colnames(mtcars)
fit<-lmFSA(formula="mpg~hp+wt",data=mtcars,fixvar="hp",
quad=FALSE,m=2,numrs=10,cores=1)
print(fit) #print formulas of fitted models
summary(fit) #review
```

```
#Logistic Regression Example
dat<-read.csv("http://tinyurl.com/zq71775",header = FALSE)
colnames(dat)<-c("Class","Age","Sex","Sterioid","Antivirals",
"Fatigue","Malaise","Anorexia","Liver Big", "Liver Firm",
"Spleen Palpable","Spiders", "Ascites","Varices","Bilirubin",
"Alk Phosphate","Sgot","Albumin","Protime","Histology")
dat<-as.matrix(dat)
dat[which(dat=="?")]=NA
dat<-data.frame(dat)
dat[,c(2,15,16,17,18,19)]<-lapply(X = dat[,c(2,15,16,17,18,19)],
as.numeric)
colnames(dat)
fit<-glmFSA(formula="Class~Age+Sgot*Albumin",data=dat,
fixvar="Age",quad=FALSE,m=2,
numrs=10,family="binomial",cores=1)
```

3.2.3 Availability

Currently, **rFSA** (version 0.1.0) is available to download from the Comprehensive **R** Archive Network at `https://cran.r-project.org/web/packages/rFSA`. To install the newest beta version of **rFSA**, first install the **devtools** package in **R**, then run the following command:

```r
devtools::install_github("joshuawlambert/rFSA")
```

.

3.2.4 Shiny App

An easy to use Shiny application has been built to facilitate the basic functions of the package. We believe this Application will serve an important role for those unfamiliar with R, but who would still like to explore subsets of large datasets or possible interactions that exist. This application allows users to use their own data and select function presets via radio buttons and drop down boxes on a server hosted by the University of Kentucky. The application on our Shiny Server is hosted at `https://shiny.as.uky.edu/mcfsa/`.

3.3 Comparisons to Other Packages

R packages currently exist to find best subsets and also explore pairwise interactions. Some of the most popular packages in current use are the **leaps** [92] package and **glmulti** [98] . These packages utilize criterion functions such as r-squared, adjusted r-squared, mallows Cp, residual sum of squares, BIC, AIC and others to find the best subset of predictor variables for a given response variable.

The **leaps** [92] package uses exhaustive search, forward and backward selection, and a sequential replacement algorithm to find the best subset of predictors in a model. The sequential replacement algorithm used in the leaps package is another variation of the FSA which [91] spoke of. While the leaps package is flexible and robust, currently, there is no way to include interaction terms. The **glmulti** [98] package is capable of adding pairwise interaction terms and uses exhaustive search, or a genetic algorithm. Both packages, address large datasets in different ways. **leaps** has a ”really.big” option which must be set to TRUE if you wish to perform exhaustive search on more than 50 variables. **Leaps**’ sequential replacement algorithm and **glmulti**’s genetic algorithm seek to provide a speedy option when there are many variables.

Timing Comparisons

A simulation was done to compare the timing of the subset selection methods in **leaps**, **glmulti**, and **rFSA**. Simulations were conducted over a grid of

$$p = (10, 20, 26, 50, 100, 150, 200, 250, 300, 500, 1000, 2000, 5000),$$

and $N = 250$ for twenty five random datasets each. A continuous response was randomly generated from a standard normal distribution.

Half of the predictors were randomly generated from a standard normal distribution and the other half were randomly generated from a Bernoulli distribution with $P(X = 1) = 0.50$. Each method was performed on all datasets and separately timed. Simulations were run in **R** version 3.1.2 on a Windows 7 machine with Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz with 24.00 GB of memory. The following commands were used in packages **leaps**(version 2.9), **glmulti**(version 1.0.7), and **rFSA**(version 0.1.0) respectively.

```r
regsubsets(x=...,y=...,nbest=1,nvmax=2,
really.big=T, method="exhaustive")

bestglm(Xy=..., family=gaussian, IC="AIC",
method="exhaustive",nvmax=2)

lmFSA(X1~1,data=...,interactions=F, m=2,
numrs=1,criterion=AIC,minimax="min)
```

Figure 1 compares the run time (seconds) for these commands over 25 simulations and $p = (10, 20, 26)$. **glmulti** failed to run for $p > 26$ in our tests. The

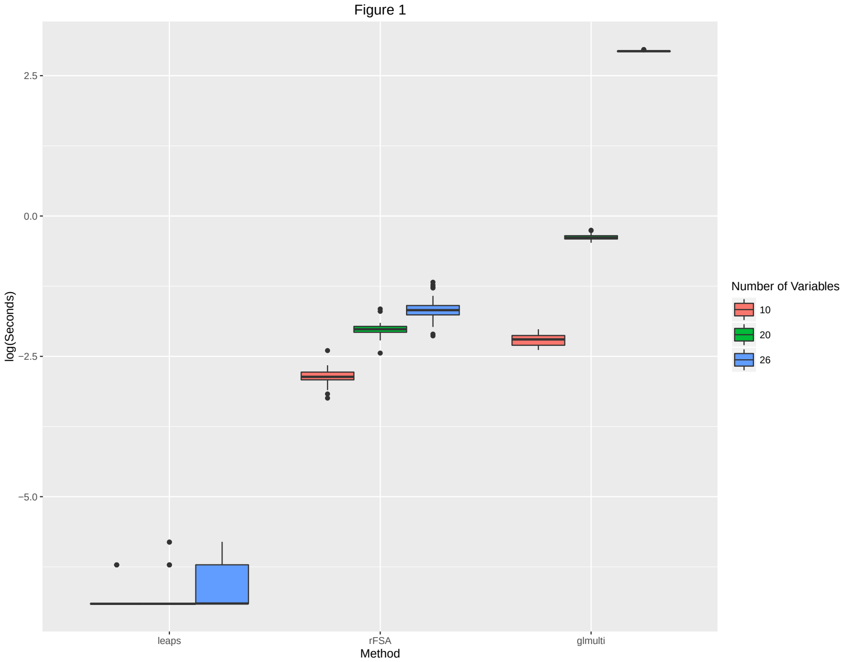


Figure 3.1: Time Comparisons of Three Methods: leaps::regsubsets, rFSA::lmFSA, and glmulti::glmulti for finding the best subset of size 2 for 25 simulations. Interactions were not considered. lmFSA times are shown for one random start, while regsubsets and glmulti times are shown for exhaustive search. Time (in Seconds) is presented on a log scale. glmulti is not presented because it failed to run for values of p greater than 26 in our simulations.

leaps::regsubsets function with the exhaustive method showed the best run time for all methods. One random start for **rFSA::lmFSA** had the next fastest run times followed by **glmulti::glmulti** using the exhaustive method. Figure 2, shows the comparison between **leaps::regsubsets** and **rFSA::lmFSA** for all values of p previously stated. While **leaps::regsubsets** is faster for all, both methods are very close when $p = 2000$.

glmulti and **rFSA** have a number of advantages over **leaps**. Both can include pairwise interactions, fit generalized linear models, and consider unique user defined criterion functions. **rFSA** is able to include even higher order interactions (example: 3-way or 4-way). Depending on the task, all three of these packages could be useful to a data scientist or statistician.

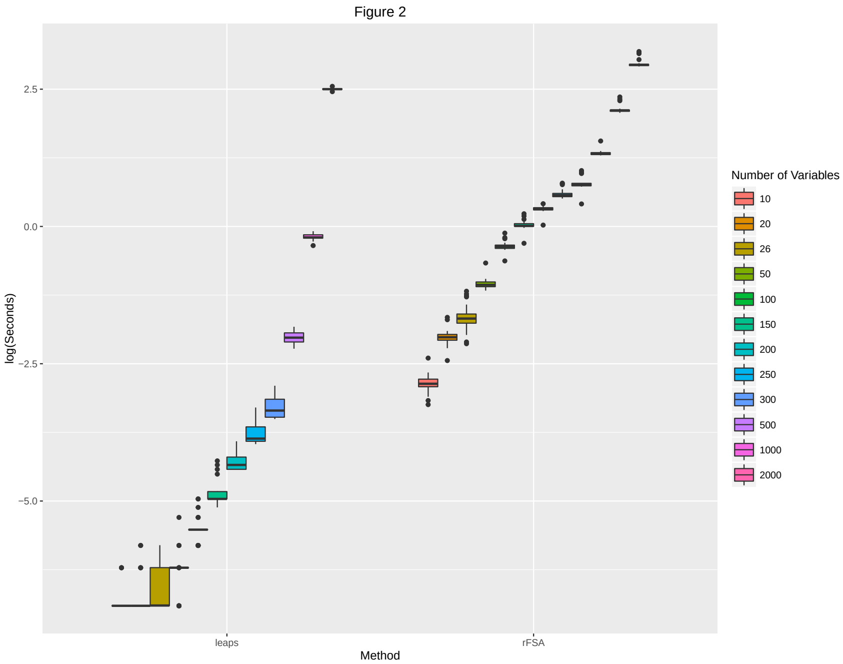


Figure 3.2: Time Comparisons of Two Methods: leaps::regsubsets, and rFSA::lmFSA for finding the best subset of size 2 with 25 simulations. Interactions were not considered. lmFSA times are for one random start, while regsubsets times are for exhaustive search. Time (in Seconds) is presented on a log scale.

While **rFSA** does not implement an exhaustive search, the optimal solution will be produced with high probability when enough random starts are completed. Running many random starts often requires fewer commutations than running an exhaustive search method. For large p, we argue that **rFSA** is a practical solution for statisticians or data scientist who wish to consider specific model forms, or generalized linear models for finding best subsets and interactions.

3.4 Example

3.4.1 Census Data Example

The data used is the publicly available 2014 Planning Database Block Group Data (PDB) from the Census Bureau at `http://www.census.gov/research/data/planning_database/2014/`. Only Kentucky Census Blocks were used and variables were removed from the dataset because they were transformations of other variables. The final dataset included 3285 observations and 67 quantitative explanatory variables and the quantitative response variable, Mail Response Rate. Descriptions of the variables can be found on the PDB documentation PDF from the website above.The final dataset used for this example can be downloaded from `https://raw.githubusercontent.com/joshuawlambert/rFSA/master/census_data_nopct.csv`.

For this dataset we wished to search for the best linear model with two main effects and their interaction. For simplicity, we chose to not fix any variables in the model. To do this we left **fixvar** equal to NULL. The response variable y (Mail Response Rate) with an intercept was the original model chosen. From this, **lmFSA** knew to use y as the response variable throughout the procedure. The criterion function, interaction p-value, was minimized at each potential swap. Exactly 50 random starts were chosen to be computed and **lmFSA** completed these 50 random starts in about one minute of run time on a Windows 7 machine with Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz with 24.00 GB of memory. R Code to reproduce these results are below.

```r
#Example 4.1
library(rFSA)
census_data_nopct <- read.csv("https://raw.githubusercontent.com/joshuawlambert/rFSA
    /master/census_data_nopct.csv")[,-1]
fit<-lmFSA(formula = y~1 ,data=census_data_nopct, fixvar = NULL, quad = F, m = 2,
    numrs = 50, cores = 1, interactions = T, criterion =int.p.val, minmax = "min")
print(fit) #summary of solutions found
summary(fit) #list of summaries from each lm fit
plot(fit) #diagnostic plots
```

Out of the 50 random starts, there were three unique feasible solutions. The solutions showed potentially interesting interactions between MrdCple_Fmly_HHD_CEN_2010 and

Pop_18_24_CEN_2010, avg_Agg_HH_INC_ACS_08_12 and avg_Agg_House_Value_ACS, and Mobile_Homes_ACS_08_12 and Tot_Vacant_Units_CEN_2010. All three of the interaction $p - values$ passed the Bonferroni cutoff criterion of 0.00045, ie $0.05/\binom{67}{2}$, at $6.19 \times 10^{-}38$, $1.68 \times 10^{-}43$, and $7.26 \times 10^{-}19$ respectively.

Variable definitions include: MrdCple_Fmly_HHD_CEN_2010: Number of 2010 Census households in which the householder and his or her spouse are listed as members of the same household; does not include same-sex married couples. Pop_18_24_CEN_2010: Number of persons ages 18 to 24 as of April 1, 2010. avg_Agg_HH_INC_ACS_08_12: Average aggregate household income. avg_Agg_House_Value_ACS_08_12: Average aggregate House value (in dollars). Mobile_Homes_ACS_08_12: Number of Mobile Homes. Tot_Vacant_Units_CEN_2010: Total vacant Housing Units in the 2010 Census.

After finding these feasible solutions, it is often useful to see a summary of their model fit. To do this, simply type **summary(fit)**, where **fit** is a FSA object. This will return a list of summaries from the model fits found in the FSA object **fit** including the original fit specified by the user. Assessing the fit of each feasible solutions is also useful, and can be done with the **plot(fit)** on the FSA object **fit**. Here, diagnostic plots are shown for the original fit and all other Feasible Solutions. Each solution should be considered in a heuristic and practical manner. Interpretation of the interaction should be considered before including it in the final model.

These three interactions were particularly interesting in understanding the types of results the algorithm gives. One might expect that a larger average household income would tend to be positively correlated with the average house value in a Census block. So, in this case, the presence of possible multicollinearity may provide an additional explanation of relationships present in the data. Another interaction that was found, MrdCple_Fmly_HHD_CEN_2010 and Pop_18_24_CEN_2010, may be more meaningful in the prediction of mail response rate (y). It seems reasonable to suspect that Census blocks with many married couples and fewer younger adults would be more likely to respond to a mailed items than a Census blocks with fewer married couples and many younger adults. Because this interaction is both interpretable and statistically significant, we have considerable evidence to justify its inclusion into the final model.

In summary, analyzing this census data using **rFSA** has provided multiple possible interactions that could be included, as well as interesting explanations existing in the data. We suggest, as with all data analysis, to first explore univariate relationships in the data. Then upon arriving at a model with known single effects (either from forward, backward, or exhaustive selection), fix those in the model through the parameter **fixvar** in **lmFSA** or **glmFSA** and proceed with exploring higher order interactions.

3.4.2 Ethical Considerations

Like any exploratory activity, one can argue that FSA is simply "data fishing." Backstops can be placed on an algorithm or method to deter users from misinterpreting or misusing. One example of such a backstop that is currently in place is by only allowing the algorithm to consider interactions that have sufficient sample size. If too many data values are missing for the variables involved in the interaction under consideration, then the algorithm excludes that interaction from those to check. Another possibility, which currently is not in place in rFSA, is allowing the user the ability to use a validation set to check identified interactions. Appropriate vignettes, tutorials, and other documentation are also essential in making sure the user understands the output and its meaning.

3.5 Conclusion and Future Work for rFSA

In this paper we have demonstrated the implementation of a complex algorithm originally proposed by [91] and [90] in an **R** package that is freely available on CRAN, **rFSA**. Additionally, we provide users with a convenient graphical user interface in the form of a Shiny Application. The application of FSA on a census data set showed the versatility and computational efficiency of the algorithm.

In terms of identification of interactions, especially in the case of data sets with a large number of explanatory variables, FSA provides an implementation of a data analysis technique that is computationally feasible while producing interpretable models and model coefficients. FSA can be applied to both quantitative and categorical response variables within standard statistical models, and this versatility remains within the R package, **rFSA**.

In order to continue improving FSA as a technique, and to create even more flexibility within the R package, we plan to include more criterion functions, and make an off line version of the Shiny App for users with secure data. In addition, considering methods to further improve the speed of the algorithm, and better utilization of the **parallel** package will continue to improve the usability of **rFSA** in analyzing large data sets. Through its versatility and flexibility, FSA provides an alternative algorithm for model selection that allows users to find subsets and interaction effects in a variety of data sets that are statistically optimal. Improved selection of such models may lead to models with improved predictive power that can in turn help illuminate relationships in large data sets.

3.6 More about the Shiny Application for rFSA

This chapter presents a Shiny application for finding interactions in large datasets. The rFSA package functions lmFSA and glmFSA are implemented in a easy to use web interface. This application can help improve existing models used in bioinformatics, health care, or other fields which have yet to explore interactions because of the size of their datasets or because they are unfamilar with statistical programming. It utilizes multi-core processing and allows users to upload their own data. It can be accessed from `shinyfsa.org` via Chrome and Firefox, but not Internet Explorer.

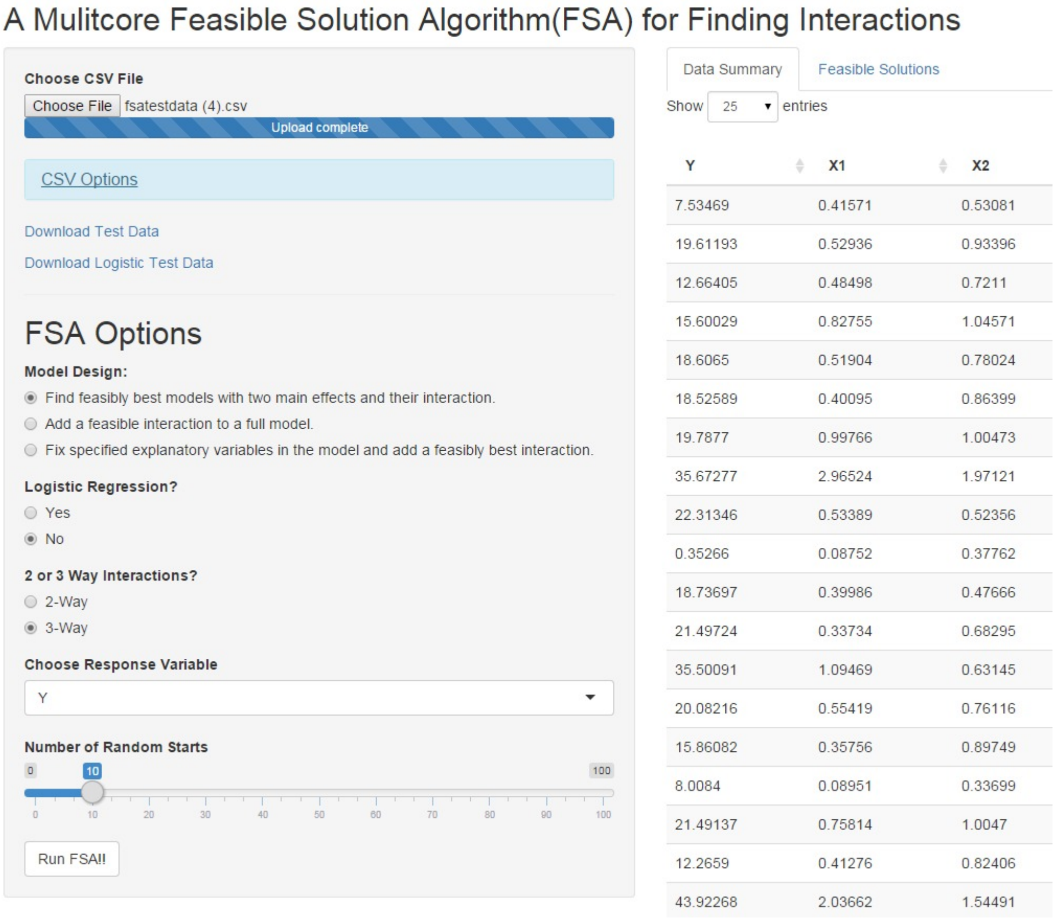
Often times, new and novel statistical methods are put forth in the form of an R package and deployed via R's Comprehensive R Archive Network (CRAN) for users to freely access. These packages can be difficult to learn and implement for individuals who have little experience with R, or who are unfamiliar with the writers syntax and/or logic. Also, computing limitations can be an issue as the datasets becomes larger. The need for a user interface that incorporates multi-core processing and cutting edge computing power is important for new statistical concepts to be adopted and used regularly. Shiny, a package in R, allows users to easily design applications that can be deployed locally or over the Internet (via Shiny's Free 'Shiny Server' software.) Shiny apps, can be deployed on a powerful server to designated or undesignated users for free. These apps can allow for users to upload their own data, and can be configured to be secure so users with protected data can have access to the same computing power and user friendly interface as those without secure data. These apps can help make statistical analysis more approachable and can help facilitate interest in new statistical methods across disciplines and fields.

3.6.1 FSA Shiny Application (FSAA)

Using a web browsers, users should first navigate to `shinyfsa.org`. Upon arrival (Figure 1), users are greeted with user options on the left and data summary/results on the right. First, users should upload data of their own, or download one of the test data sets provided below the "Browse..." button. The FSA application (**FSAA**) allows users to upload their own data in a CSV (comma separated values) format, with one subject/unit per row and one variable per column. Extraneous information should be excluded from the CSV files before upload. Users should see the test data sets for examples of acceptable CSV formats. Users with Categorical Variables, should make sure that these variables are in quotes before upload. This will assure the user that these variables are being modeled as categorical and not continuous.

The data that are uploaded are immediately deleted once the application is closed. Users should not upload data that requires protected or encrypted web protocol. Users with secure data should only use the rFSA package on a local personal machine or server. Users can click the "Browse..." button at the top left of the application to browse their personal machine for the CSV file of their liking. Once the dataset is selected, users should click the "Open" button to upload their data. Once the upload has completed successfully, the "Data Summary" tab on the right will show some of the data from within the dataset. If no column names are included, please un-select the header option under "CSV Options", and the application will name the variables $V1 - V_{p+1}$.

Figure 3.3: FSA Shiny Application: Data Summary



A Mulitcore Feasible Solution Algorithm(FSA) for Finding Interactions

Choose CSV File

Choose File fsatestdata (4).csv

Upload complete

CSV Options

Download Test Data

Download Logistic Test Data

FSA Options

Model Design:

● Find feasibly best models with two main effects and their interaction.

○ Add a feasible interaction to a full model.

○ Fix specified explanatory variables in the model and add a feasibly best interaction.

Logistic Regression?

○ Yes

● No

2 or 3 Way Interactions?

○ 2-Way

● 3-Way

Choose Response Variable

Y

Number of Random Starts

0 10 100

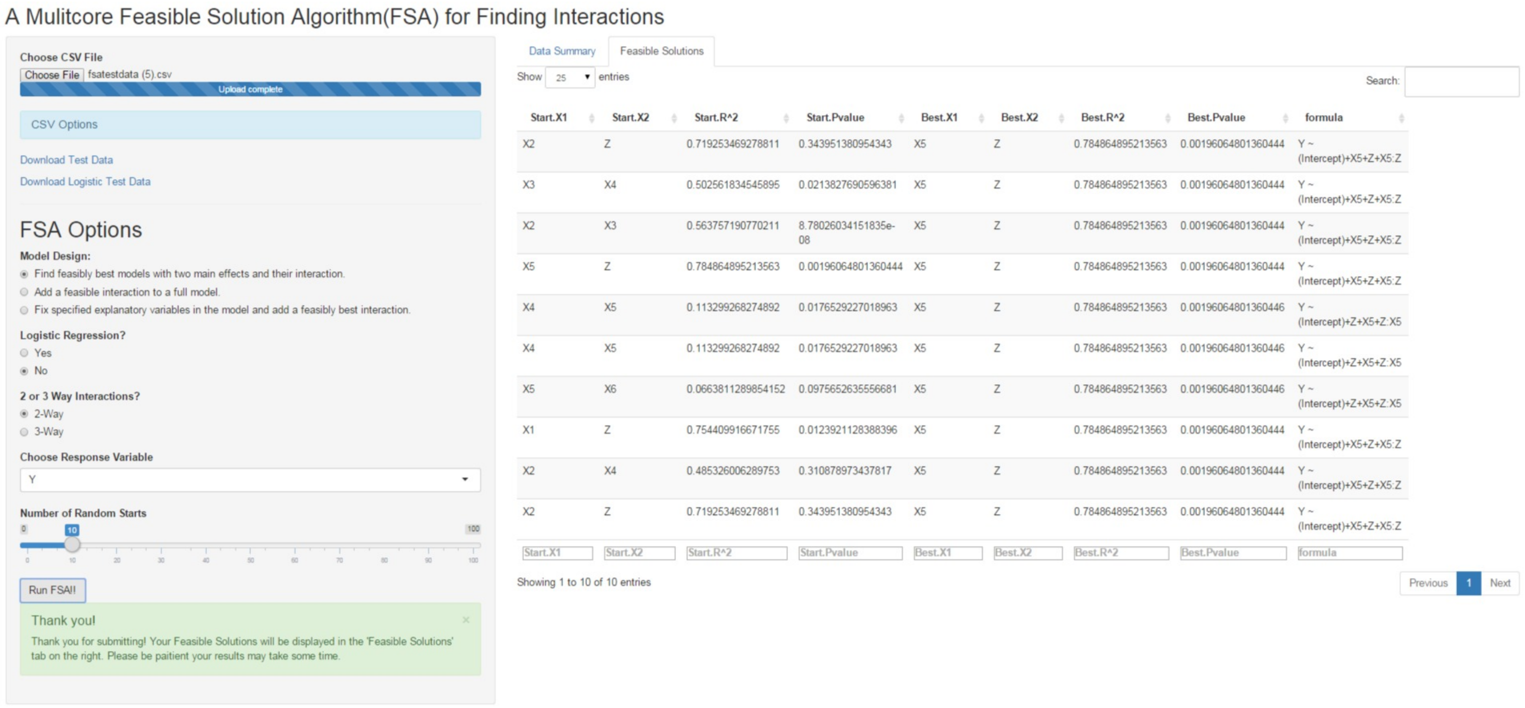
Run FSA!!

Data Summary | Feasible Solutions

Show 25 entries

Y	X1	X2
7.53469	0.41571	0.53081
19.61193	0.52936	0.93396
12.66405	0.48498	0.7211
15.60029	0.82755	1.04571
18.6065	0.51904	0.78024
18.52589	0.40095	0.86399
19.7877	0.99766	1.00473
35.67277	2.96524	1.97121
22.31346	0.53389	0.52356
0.35266	0.08752	0.37762
18.73697	0.39986	0.47666
21.49724	0.33734	0.68295
35.50091	1.09469	0.63145
20.08216	0.55419	0.76116
15.86082	0.35756	0.89749
8.0084	0.08951	0.33699
21.49137	0.75814	1.0047
12.2659	0.41276	0.82406
43.92268	2.03662	1.54491

Figure 3.4: FSA Shiny Application: Feasible Solutions



A Mulitcore Feasible Solution Algorithm(FSA) for Finding Interactions

Choose CSV File

Choose File fsatestdata (5).csv

Upload complete

CSV Options

Download Test Data

Download Logistic Test Data

FSA Options

Model Design:

- ● Find feasibly best models with two main effects and their interaction.
- ○ Add a feasible interaction to a full model.
- ○ Fix specified explanatory variables in the model and add a feasibly best interaction.

Logistic Regression?

- ○ Yes
- ● No

2 or 3 Way Interactions?

- ● 2-Way
- ○ 3-Way

Choose Response Variable

Y

Number of Random Starts

10

Run FSA!!

Thank you!

Thank you for submitting! Your Feasible Solutions will be displayed in the 'Feasible Solutions' tab on the right. Please be paitient your results may take some time.

Start.X1	Start.X2	Start.R^2	Start.Pvalue	Best.X1	Best.X2	Best.R^2	Best.Pvalue	formula
X2	Z	0.719253469278811	0.343951380954343	X5	Z	0.784864895213563	0.00196064801360444	Y ~ (Intercept)+X5+Z+X5:Z
X3	X4	0.502561834545895	0.0213827690596381	X5	Z	0.784864895213563	0.00196064801360444	Y ~ (Intercept)+X5+Z+X5:Z
X2	X3	0.563757190770211	8.78026034151835e-08	X5	Z	0.784864895213563	0.00196064801360444	Y ~ (Intercept)+X5+Z+X5:Z
X5	Z	0.784864895213563	0.00196064801360444	X5	Z	0.784864895213563	0.00196064801360444	Y ~ (Intercept)+X5+Z+X5:Z
X4	X5	0.113299268274892	0.0176529227018963	X5	Z	0.784864895213563	0.00196064801360446	Y ~ (Intercept)+Z+X5+Z:X5
X4	X5	0.113299268274892	0.0176529227018963	X5	Z	0.784864895213563	0.00196064801360446	Y ~ (Intercept)+Z+X5+Z:X5
X5	X6	0.0663811289854152	0.0975652635556681	X5	Z	0.784864895213563	0.00196064801360446	Y ~ (Intercept)+Z+X5+Z:X5
X1	Z	0.754409916671755	0.0123921128388396	X5	Z	0.784864895213563	0.00196064801360444	Y ~ (Intercept)+X5+Z+X5:Z
X2	X4	0.485326006289753	0.310878973437817	X5	Z	0.784864895213563	0.00196064801360444	Y ~ (Intercept)+X5+Z+X5:Z
X2	Z	0.719253469278811	0.343951380954343	X5	Z	0.784864895213563	0.00196064801360444	Y ~ (Intercept)+X5+Z+X5:Z

Showing 1 to 10 of 10 entries

After the data are uploaded correctly, then users can sift through their data by sorting by a certain column or searching for certain variables or variable values. Please note that searching, or sorting does not change the original data or the data that will be run in the analysis.

Next, users can choose their Model Design. If the response variable is binary, then they will want to select ”Yes”, under the logistic regression category. The equation to find feasibly best models with two main effects and their interaction is: $y = \beta_0 + \beta_1 x_a + \beta_2 x_b + \beta_3 x_a x_b$. The equation to add a feasible interaction to a full model will be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \beta_3 x_a x_b$. And finally, the equation to fix specified explanatory variables in the model and add a feasibly best interaction, for example fixing x_2 and x_{25}, is $y = \beta_0 + \beta_1 x_2 + \beta_2 x_{25} + \beta_3 x_a + \beta_4 x_b + \beta_5 x_a x_b$. After selecting your model type, users need to determine the number of random starts to do. It is preset to 10, but users can select up to 100. If users select 100, you may experience some processing delays depending on the number of p variables. Lastly, users should select whether they want to look for two-way or three-way interactions.

Once the user has selected the options appropriate for their data, they are then ready to click the ”Run FSA!!”, button. After clicking (figure 2), a thank you message will appear below the ”Run FSA!” button and **FSAA** will switch from the ”Data Summary”, tab to the ”Feasible Solutions”, tab. This tab will be blank until the **FSAA** has completed. Any error messages will display here as well. If you experience an error, please explore your data in another statistical software first, and make sure that non of the explanatory variables are perfectly related with one another and/or remove any predictors that have lots of missing values. When the **FSAA** is completed, each random start will be shown on a separate row. The random start will be shown in the first two or three columns with the starting R^2 and starting interaction $p - value$ in the following columns. Next, the feasible solution that the algorithm converge to is displayed. The corresponding R^2 and $p - value$ for the main effect of the interaction for that specific feasible solution is next. Lastly, the feasible solution model is displayed so users can check what model is being fit. The ‘Feasible Solutions’ tab is also searchable and sortable based on the preference of the user.

Conclusion and Future Work for FSAA

In this paper we have demonstrated the implementation of a complex algorithm into a simple to use web application. Writing Shiny applications is easy to learn and there are a plethora of examples and resources available on the Internet to learn from and practice. We hope to encourage statisticians and data analysts to start extending their R packages into Shiny applications that are easy to use and are understandable to a wide variety of fields. The flexibility and deploy-ability of Shiny applications allow authors to quickly update and enhance their apps for new features and improvements.

We plan to improve FSAA so it can run many different instances of the feasible solution algorithm and allow users to do simple data manipulation and analysis. More options for FSAA, such as maximizing via AIC, BIC, AUC, and whether to include quadratic terms are features we hope to include soon. Increasing the number of cores and RAM available to the user is another future improvement we would like to make, this way users can upload larger datasets and increase the number of random starts. Increasing the level or security and providing account based access to the application is also a feature we believe could be useful for users with sensitive data, and can be implemented with Shiny Server Pro (Paid).

Chapter 4 Exploring Statistical Interactions associated with Periodontal Disease in NHANES 1999-2004 Data

Periodontitis is a complex and multi-faceted disease involving smoking, sex, age, Socio Economic Status (SES), race, as well as others [30][40] [41] [15] [33] [99] [100]. Interrelationships of these complexities can be investigated via statistical interactions. Previously, there has been a lack of model-based approaches to identify higher order interactions. In this chapter, higher order interactions related to periodontal disease are explored using a Feasible Solution Algorithm (FSA) via the R package: rFSA [101]. Periodontal status (either present or absent) calculated from NHANES survey data, from 1999-2004, will be analyzed using Survey Weighted Logistic Regression and FSA. The FSA seeks to find interactions that are statistically optimal in the sense that no one swap to any of the variables included in the three way interaction will improve the underlying criterion function (in this case, interaction p-value). These interactions are further investigated for their consistency with existing literature as well as their statistical validly through tables and figures. Furthermore, discussion of the results in light of periodontal epidemiology is included.

4.1 Introduction

4.1.1 Periodontal Disease

Periodontal disease is an inflammatory process which encompasses an array of clinical features such as receding gums, loss of supporting alveolar bones, and eventually tooth loss. Periodontal disease is considered a complex disease and the factors which attribute to its onset are still greatly misunderstood. While bacterial pathogens are necessary for disease development, it is the host immune response (either over-active or under-active) that determines the extent of the tissue damage and susceptibility to periodontal disease [102]. The host response leads to the release of cytokines. These cytokines have tissue and bone destroying properties and are believed to be the primary cause of the clinical outcomes (bone loss, gum recession, ect). [103]

Recently, research has focused on identifying subgroups of susceptible individuals to periodontal disease. This has only highlighted the multi-faceted nature of the disease. Environmental factors and genetic predispositions together play a role in the disease risk profile. There is an increased prevalence of periodontitis with age. Men and ethnic minorities are at the greatest risk for periodontal disease [30]. Hispanic-Americans, Non-Hispanic Blacks, and Non-Hispanic Asian Americans all have periodontitis at levels greater than 50% [15]. Individuals with low socioeconomic status had twice the prevalence of periodontal disease compared to those with high socioeconomic status [15]. Smoking tobacco is the strongest modifiable risk factor identified for periodontitis [40] [41]; the severity of periodontitis increases with years of smoking[42].

4.1.2 Toxins and Dietary Nutrients

Toxins, such as lead and mercury, have previously been studied in terms of the pathophysiology of periodontal disease. Lead is a potent neurotoxin and contributes to the onset of many diseases [104]. Lead exposure is associated with neuropathy, increased blood pressure, altered reproductive function, and bone abnormalities [104]. Previously, a positive association was found in the NHANES III dataset (1988-1994) with lead levels and prevalance of periodontal disease. Both men and women showed a positive association with periodontal disease and blood lead serum levels [49]. A Korean National Health and Nutritional Examination Survey (KNHANES) found a positive association between blood lead serum levels and periodontitis for both females and non-smokers [105]. Lead has been shown to negatively affect bone health in animal and human studies [106] [107] [108] [108]. Interestingly, the immune host response is also modulated by lead exposure. Animal studies have demonstrated a dysregulation and inhibition of the immune response in those rats chronically exposed to lead [109] [110] [111]. Two previous studies have examined the association between mercury and periodontitis. The first study was based on a dataset that collected samples from two different Korean cities. Han et al, found that mercury exposure, assessed by mercury within hair strands, was associated with periodontitis in that sub-population of Korea [112]. The second study examined a sample of the Korean population from the KNHANES dataset (mentioned in Result 3 section) [113]. A positive association between mercury-inorganic serum levels and periodontitis was reported. Other pollutants have been previous linked to low bone density.

Alpha carotene is one of the caratenoids (alpha-carotene, beta-carotene,crytoxanthin, lutein, lycopene, and zeanxantin). Carotenoids are considered antioxidants. The balance between antioxidants and free radicals keeps a biological system free of inflammation. When this balance is disrupted, oxidative stress occurs and disease follows [114]. Since inflammation is implicated as the main cause of periodontal disease, researchers have focused on finding antioxidants that are protective towards periodontitis. One study examined periodontal health and the association with carotenoids, retinol, and Vitamin E in a group of men aged 60-70 [115]. Of particular interest, low levels of alpha and beta-carotene were significantly associated with periodontitis. In the adjusted model, older men with higher levels of alpha and beta carotene had lower estimated odds of periodontitis.

4.1.3 Methods for Identifying Interactions

Interactions can be searched for on an individual level by statisticians . This is usually based on previous knowledge of the data, or expertise of the primary investigator. These processes can be tedious and can lead to interactions being ignored or missed due to the large number of interactions to check. Exhaustively searching for the optimal model that includes interactions is not always computationally possible or feasible.

Analytical methods exist to investigate potential interactions. Interactions can be explored via Classification and Regression Tree (CART) models. Random Forest [94] and Boosted Trees [95] methods fit many CART models. In these models, consistencies in which variables are branched on can give insights into potential modifiers. Bayesian methods also exist for exploring interactions [96] [116].

Recently, a new algorithm for exploring interactions has been formalized to use existing statistical models and search the data using the Feasible Solution Algorithm (FSA) for a set of semi-optimal higher order interactions for consideration [101].

CART models and Random Forest models have been used for making better periodontal classifiers [117], finding periodontal prognostic indicators [118], identifying pathogen and host-response markers related to periodontal disease [119], and environmental factors related to periodontal disease [120]. While these CART and Random Forest models have demonstrated interesting results, model-based statistical interactions related to periodontal disease has not been fully explored [121]. By identifying model-based interactions, biases , misinterpretation, and incorrect public health interventions are avoided[122]. Model based approaches are capable of adjusting for known risk factors which is important when looking for small effects that may be masked. Also, statistical models exist to appropriately account for survey weights. Continuous variable estimation is also possible with statistical models, whereas methods such CART does not.

This paper explored the interrelated role of environmental factors, dietary nutrients, and demographic variables on periodontal diseases by identifying model based three-way interactions that are associated with the prevalence of periodontal disease. Because FSA can identify model based interactions and handle a large number of variables, it will be used to explore interactions in the data. The three way interactions under consideration will be limited, in that they must have at least one of the five identified risk factors related to periodontal disease that are also in the NHANES data.

4.2 Materials and Methods

4.2.1 Data and Variables

Data

Partial mouth periodontal examination data from three cohorts of NHANES data from 1999-2004 were extracted and combined with environmental and demographic information for the same time period. All data were download freely from NHANES website. Before exclusion, there were 11,041 participants. There were 10,277 NHANES participants who had at least 16 teeth, were older than 18, had partial periodontal examination measurements, and a non missing smoking status.

Demographic Variables

There were five variables targeted as demographic covariates. These variables are age (RIDAGEYR), sex (RIAGENDR), race (RIDRETH1), family income to poverty (INDFMPIR), and smoking status (SMQ020 and SMQ040). Race/Hispanic origin was summarized into five categories: Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, and Other Race. Family Income to poverty level will act as an estimate of Socio Economic Status. For smoking status, participants who answered that they had not smoked more than 100 cigarettes and were not currently smoking were defined as former smokers. Non-smokers were defined as those who had reported smoking more than 100 cigarettes and were not currently smoking. Smokers were defined as those who had reported smoking more than 100 cigarettes and were either smoking "Every day" or "Some days".

Clinical Variables

To sample teeth and sites, the NHANES 1999-2004 surveys followed a partial-mouth periodontal examination (PMPE) protocol. This protocol, measures pocket depth, attachment loss, and bleed on probing for 2 (1999-2000) to 3 (2001-2004) sites per tooth from two randomly selected quadrants. The Centers for Disease Control (CDC) and National Institute of Dental and Craniofacial Research (NIDCR) guidelines indicate a periodontal tooth is defined by having a clinical attachment loss (CAL) $\geq 3mm$ and a periodontal pocket $\geq 4mm$ (OHDLAM, OHDLAS, OHDPCM, OHDPCS, OHDPD). Subsequently, a subject with periodontitis was defined as someone who had at least one periodontal tooth following the NIDCR guidelines.

Environmental variables

Environmental variables considered must have been present in at least one of the three data cohorts and had to have at least 10% of its measurements above the limit of detection threshold indicated by NHANES [120]. The final 156 environmental variables extracted for analysis, all came from one of the following categories: chemical toxicants, pollutants, allergens, bacterial/viral organisms and nutrients. Many of these variables were detected in small ranges and skewed. Therefore they all were processed by taking the log(natural) and then standardized. There are 34 PCB variables in the dataset. Many polychlorinated biphenyl (PCB) accumulate in the body and other organisms over time [123][124]. Therefore, to estimate the cumulative effect of all 34 PCB's in the NHANES data, all 34 PCB variables were first added then standardized.

Potential Confounders and Effect Modifiers

Sex, age, Socio Economic Status (SES), race, and smoking status have all been identified previously as risk factors or confounders for periodontal disease [15] [30] [29] [33][42]. Smoking, and Sex have been previously mentioned as potential modifiers related to periodontal disease[121]. SES, age, and race could be considered risk factors whose effects could be modified by other covariates or known risk factors.

4.2.2 Statistical Analysis

The focus of this statistical analysis was to identify three way interactions that could be included in a survey weighted logistic regression model while adjusting for known risk factors, and confounders. Because of the model based approach, the FSA was used to identify potential interactions. An emphasis was placed on three way interactions that consisted of at least one of the known risk factors or confounders that were available in the data. By doing this, we are able to identify subgroups of patients based on these known factors as well as uncover new relationships with other risk factors and environmental variables.

The FSA is an algorithm that is used along side a specific statistical model regression method for identifying new variables, quadratic terms, or interactions that could be added to the model. First, the appropriate statistical model must be decided on. For our data, survey weighted logistic regression is the appropriate statistical model to analyze these data.

When the response is binary (Periodontitis or No Periodontitis), and there are survey data with appropriate weights available, a Survey Weighted Logistic Regression is used to obtain parameter estimates and standard errors for statistical inference. This method minimizes the bias that can exist from the way the samples are collected. It does this by weighting the samples to reflect the intended population. By doing this, standard error estimates for the regression coefficients are correctly obtained. To correctly adjust for these issues and obtain correct standard error estimates, a set of survey weights and the groupings of how the variables are clustered is needed. NHANES 1999-2004 have both of these (Using NHANES Weights, and Making Weights). Using R [125] and the *Survey* [126] R package, survey weighted logistic regression can be performed. Providing the *svyglm* function in the *Survey* package a model formula, survey design, data, and the appropriate model family a survey weighted logistic regression model can be fit.

Once a statistical regression method is identified, the FSA will checks combinations of variables for their significance in the model via a criterion function chosen by the user. For three way interaction identification, FSA works by first adding a random three way interaction and checking its significance in the model. Then, the FSA will check variations of the randomly chosen three way interaction by exchanging two variables in the current interaction for the other possible two variable combinations. Each combination is checked for its significance in the model. The most significant combination is then chosen (swapped to) as the next starting place and the process is repeated. Usually after swapping 2-3 times, a combination is found that cannot be improved upon by exchanging only one of the variables in the currently optimal three way interaction. Many iterations (random starts) of this process will yield a set of feasible solutions. Some random starts will give the same feasible solutions as others, while others may be the only random starts that give that specific feasible solution. A more technical overview of the FSA and rFSA can be found in Chapter 3.

Approach

Data from 1999-2004 will be analyzed using R [125] V3.4.1 and package *Survey* [126] 3.32-1. Adjusted Odds Ratios, 95% confidence intervals, and p-values were computed for each model only with the interactions main effects (Model A) and its full model with the three way interaction and its lower order terms (Model B). These two models are shown to compare how the interaction model is different then just using the main effects. Sex, Age, SES, Race, and Smoking status were identified as covariate effects that would need to be first adjusted by including them as main effects in the linear model before identifying potential interactions. These same variables are risk factors which could potentially modify the effects of environmental toxins/dietary nutrients on prevalence of periodontal disease. For that reason, at least one of these variables was required to be in the resulting interaction. Twenty random starts were performed for each analysis (100 total) and the *anova.svyglm* R function was used to extract the interaction p-value from a sequential anova table. FSA sought to find feasible solutions by way of minimizing that interaction p-value.

Each result includes a statistical table with the main effects model. This table is meant to highlight what was present without the interactions that was found. To visualize the interaction that was found a graph is included for interpretation. Two variables were chosen to dichotomize on the median of the variables for graphical purposes. The median split was chosen to try and keep group sizes as equal as possible. While this was done for graphical purposes, the variables were still modeled as continuous in the statistical model. From now on, when referring to the figure, those subjects described as having higher levels for a variable are referent to those subjects that have levels above the median, while those subjects that are described as having lower levels for a variable are referent to those subjects that have levels lower then the median.

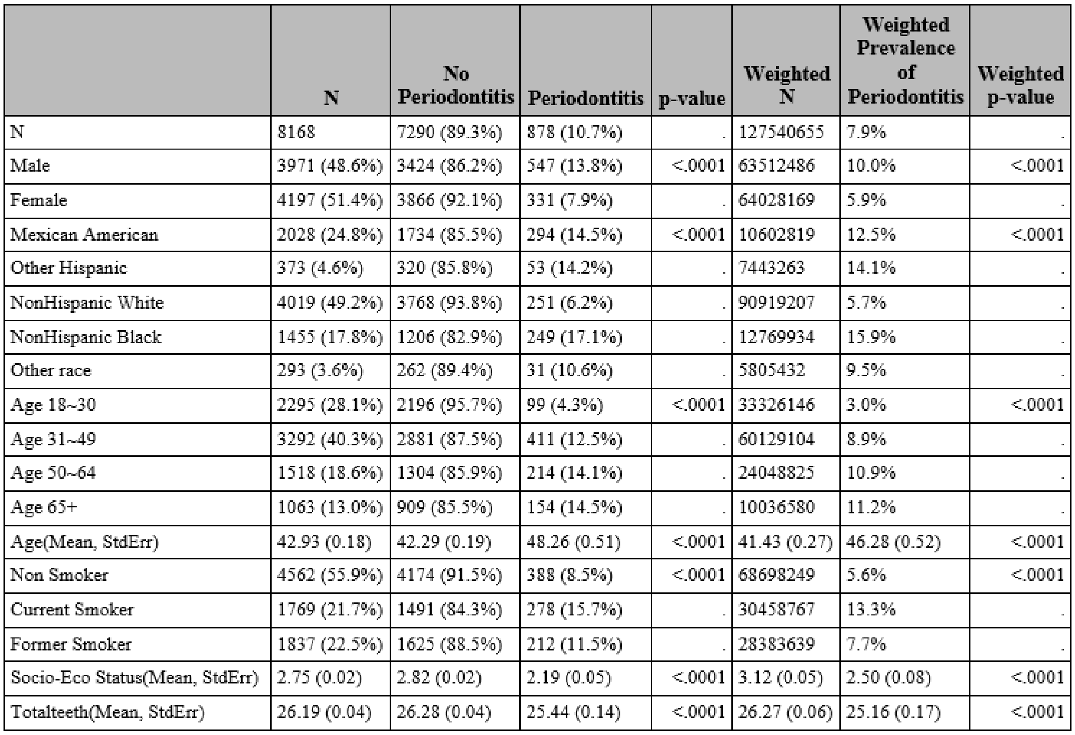
The y-axis corresponds to the predicted probability of periodontitis from the model for a given participant. This probability comes from the full model (A) that includes the three way interaction. By plotting the probability of periodontitis for an individual we are able to gain greater insight into the effect of interaction across the sample.

The following results are broken into two groups. Group 1 is the results that included just one of the five risk factors and toxins and/or dietary nutrients for the other two factors in the interaction. Group 2 is the results that included two of the five risk factors and just one toxin or dietary nutrient for the other factor in the interaction. All of the 100 random random start's solutions can be found in the appendix table A.1 of this dissertation. Variables with "zl" at the beginning of them are variables that were first log transformed and then standardized.

4.3 Results

After excluding the 3,669 subjects who had a missing values for periodontal status, sex, age, Socio Economic Status (SES), race, and smoking status, there were 8,168 subjects who were available for analysis. Missing data analysis showed that those subjects who were older, men, identified as Mexican Americans, and those who had lower SES were all more likely to have missing values for periodontal disease. Females comprised of 51.4% of the final sample. Those who identified as Non Hispanic White (49.2%), and Mexican American (24.8%) were the most represented within the sample. Those subjects who identified as Non Hispanic Black's (17.8%), Other Hispanic (4.6%), and Other race (3.6%) were the least represented in the sample. Those aged 31-49 made up the largest age demographic at 40.3%. Roughly 22% of the sample was comprised of current smokers. Finally, 878 (10.7%) of the 8,168 subjects had at least one periodontal tooth as defined by the NIDCR. The following table 4.1 is adapted from Emecen-Huja [120].

Table 4.1: Table of Basic Demographics



	N	No Periodontitis	Periodontitis	p-value	Weighted N	Weighted Prevalence of Periodontitis	Weighted p-value
N	8168	7290 (89.3%)	878 (10.7%)	.	127540655	7.9%	.
Male	3971 (48.6%)	3424 (86.2%)	547 (13.8%)	<.0001	63512486	10.0%	<.0001
Female	4197 (51.4%)	3866 (92.1%)	331 (7.9%)	.	64028169	5.9%	.
Mexican American	2028 (24.8%)	1734 (85.5%)	294 (14.5%)	<.0001	10602819	12.5%	<.0001
Other Hispanic	373 (4.6%)	320 (85.8%)	53 (14.2%)	.	7443263	14.1%	.
NonHispanic White	4019 (49.2%)	3768 (93.8%)	251 (6.2%)	.	90919207	5.7%	.
NonHispanic Black	1455 (17.8%)	1206 (82.9%)	249 (17.1%)	.	12769934	15.9%	.
Other race	293 (3.6%)	262 (89.4%)	31 (10.6%)	.	5805432	9.5%	.
Age 18~30	2295 (28.1%)	2196 (95.7%)	99 (4.3%)	<.0001	33326146	3.0%	<.0001
Age 31~49	3292 (40.3%)	2881 (87.5%)	411 (12.5%)	.	60129104	8.9%	.
Age 50~64	1518 (18.6%)	1304 (85.9%)	214 (14.1%)	.	24048825	10.9%	.
Age 65+	1063 (13.0%)	909 (85.5%)	154 (14.5%)	.	10036580	11.2%	.
Age(Mean, StdErr)	42.93 (0.18)	42.29 (0.19)	48.26 (0.51)	<.0001	41.43 (0.27)	46.28 (0.52)	<.0001
Non Smoker	4562 (55.9%)	4174 (91.5%)	388 (8.5%)	<.0001	68698249	5.6%	<.0001
Current Smoker	1769 (21.7%)	1491 (84.3%)	278 (15.7%)	.	30458767	13.3%	.
Former Smoker	1837 (22.5%)	1625 (88.5%)	212 (11.5%)	.	28383639	7.7%	.
Socio-Eco Status(Mean, StdErr)	2.75 (0.02)	2.82 (0.02)	2.19 (0.05)	<.0001	3.12 (0.05)	2.50 (0.08)	<.0001
Totalteeth(Mean, StdErr)	26.19 (0.04)	26.28 (0.04)	25.44 (0.14)	<.0001	26.27 (0.06)	25.16 (0.17)	<.0001

4.3.1 Results 1-3: One risk factor/confounder & two toxins and/or dietary nutrients

Result 1: 3-Way Interaction: Lead, g-Tocopherol (Vit E), and Age

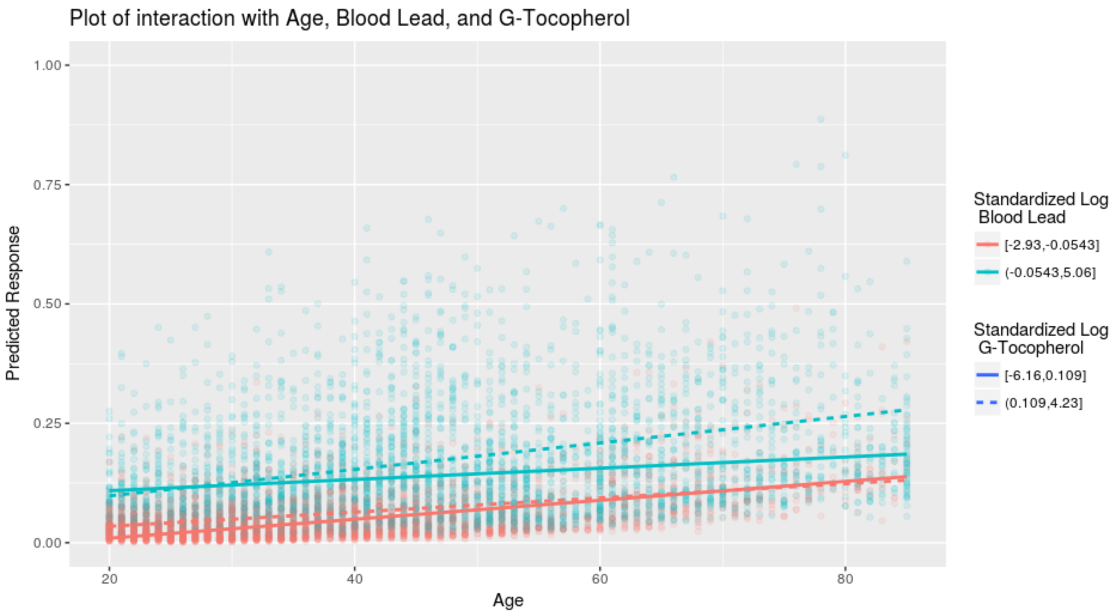
As table 4.2 shows, Age, Lead, and g-Tocopherol were all statistically significant ($p < 0.05$) for the main effects model. While these are significant, the interaction that was found between Lead, g-Tocopherol (Vit E), and Age suggest that the effects of these variables changes as the others change.

As fig. 4.1 shows, subjects predicted probability of periodontal disease increases as age increases across all interaction subgroups. Those with lower levels of lead, regardless of the level of g-Tocopherol , show similar predicted probability profiles with age. Those subjects with lead levels above the median showed different trajectories of predicted probability as age increases. Furthermore, subjects with high g-Tocopherol levels show increasingly larger predicted probability of periodontal disease as age increases.

Table 4.2: Parameter Estimates For Main Effects Model From Result 1

	OR & 95% CI	P-Value
(Intercept)	0.07 (0.04 , 0.11)	< 0.001
Age	1.03 (1.03 , 1.04)	< 0.001
SES	0.82 (0.76 , 0.88)	< 0.001
Sex	0.65 (0.54 , 0.79)	< 0.001
Race:O Hisp	1.2 (0.75 , 1.93)	0.461
Race:White	0.49 (0.34 , 0.7)	0.001
Race:Black	1.43 (1 , 2.04)	0.058
Race:Other	0.98 (0.57 , 1.67)	0.938
Smoking:Current Smoker	2.19 (1.65 , 2.89)	< 0.001
Smoking:Former Smoker	1.03 (0.79 , 1.35)	0.806
zl_Blood Lead	1.55 (1.36 , 1.77)	< 0.001
zl_g-Tochopherol	1.12 (1.01 , 1.24)	0.040

Figure 4.1: Lead, g-Tocopherol (Vit E), and Age



Result 2: 3-Way Interaction: Alpha Carotene, Toluene, and Age

For this result, there were no subjects in the dataset that had values of Alpha Carotene, and Toluene over the age of 60.

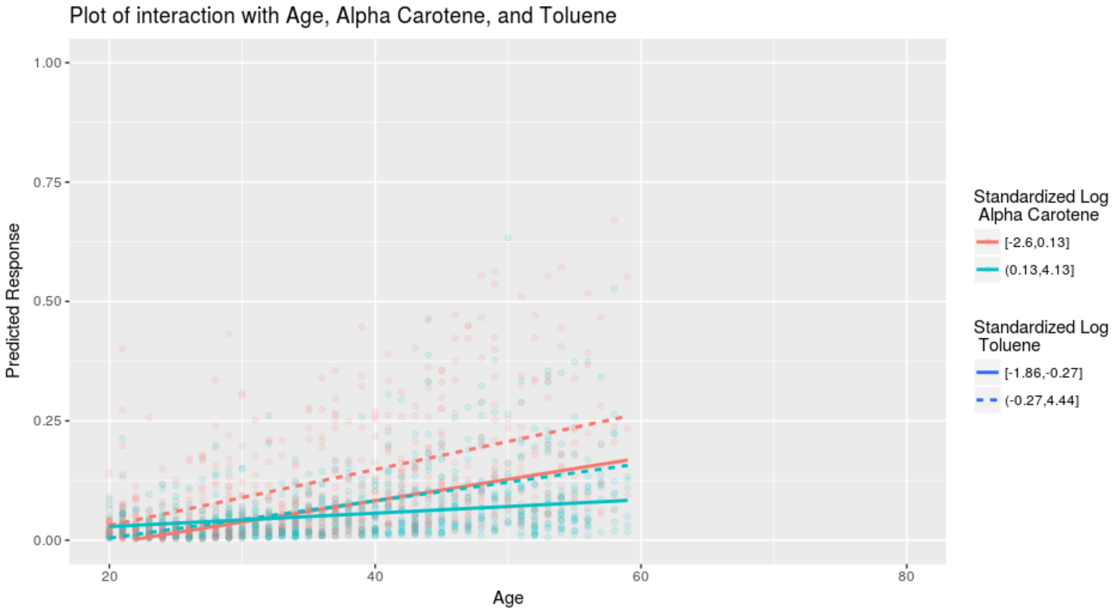
As table 4.3 shows, Alpha Carotene, and Toluene were not statistically significant ($p < 0.05$) for the main effects model. Therefore, if main effects had only been searched for then both Alpha Carotene, and Toluene would have not been added to the model. Age was statistically significant ($p < 0.05$) in the main effects model. By adding the three way interaction, and its lower order terms, the results suggests that by having high levels of Alpha Carotene, and lower levels of Toluene, individuals odds of periodontal disease with age will be its lowest. Those with low levels of Alpha Carotene and high levels of Toluene, will have the lowest odds of periodontal disease with age.

Table 4.3: Parameter Estimates For Main Effects Model From Result 2

	OR & 95% CI	P-Value
(Intercept)	0.03 (0.01 , 0.12)	< 0.001
Age	1.07 (1.05 , 1.1)	< 0.001
SES	0.79 (0.68 , 0.91)	0.005
SEX	0.53 (0.38 , 0.74)	0.001
Race:O Hisp	0.74 (0.25 , 2.2)	0.595
Race:White	0.32 (0.19 , 0.53)	< 0.001
Race:Black	0.74 (0.44 , 1.24)	0.264
Race:Other	0.73 (0.31 , 1.74)	0.490
Smoking:Current Smoker	1.92 (1.01 , 3.66)	0.062
Smoking:Former Smoker	0.85 (0.47 , 1.55)	0.600
zl_Alpha Carotene	0.82 (0.65 , 1.04)	0.124
zl_Toluene	1.2 (0.94 , 1.53)	0.154

As fig. 4.2 shows, subjects predicted probability of periodontal disease increases as age increases across all interaction subgroups. Those subjects with low values of Toluene, and high values of Alpha Carotene show the lowest trajectory of predicted probability with age for all four subgroups. Next, subjects with either low Alpha Carotene and low Toluene, or high Alpha Carotene and high Toluene showed very similar trajectories of predicted probability with age. The highest trajectory of predicted probability of periodontitis with Age were those with low Alpha Carotene and

Figure 4.2: Alpha Carotene, Toluene, and Age



high Toluene.

Result 3: 3-Way Interaction: B-cryptoxanthin, cis-Beta Carotene, and Smoking

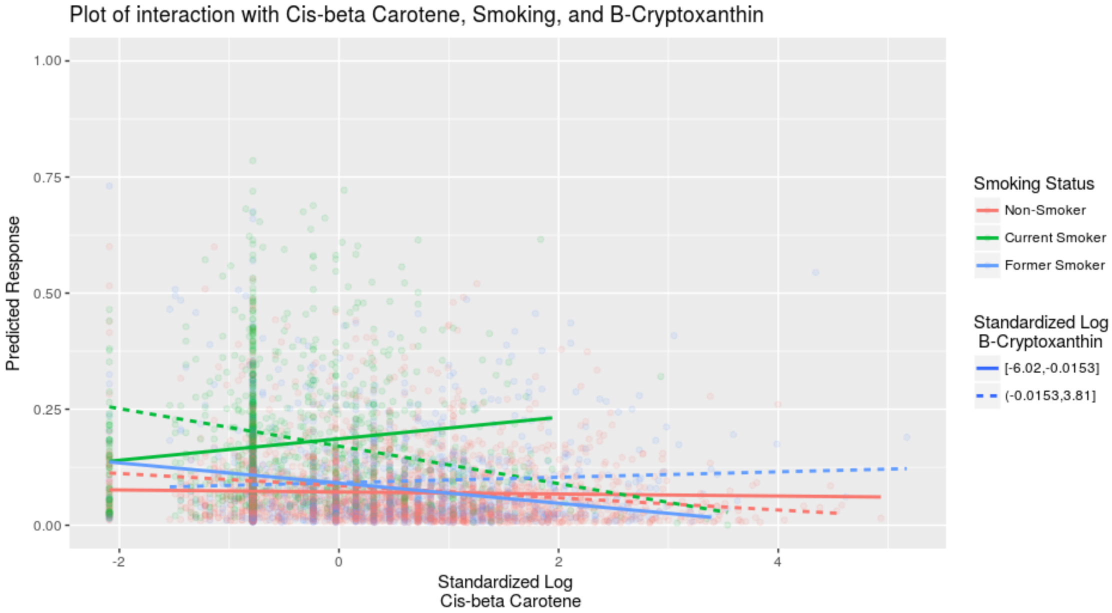
As table 4.4 shows, both B-cryptoxanthin, and cis-Beta Carotene were not statistically significant ($p < 0.05$) for the main effects model. Therefore, if main effects had only been searched for then both B-cryptoxanthin, and cis-Beta Carotene would have not been added to the model. The Current Smoker group was statistically significant ($p < 0.05$) in the main effects model.. By adding the three way interaction, and its lower order terms, this result suggests that current smokers could benefited from having high levels of both B-cryptoxanthin, and cis-Beta Carotene.

As fig. 4.3 indicates, Non and Former Smokers share a similar trajectory of predicted probability of periodontal disease which decreases as Cis-Beta Carotene increases regardless of levels of B-Cryptoxanthin. If the graph had B-Cryptoxanthin on the x-axis it would show a similar effect, which would indicate that higher levels of B-Cryptoxanthin were good regardless of the levels of cis-Beta Carotene. Current smokers, high levels of both B-Cryptoxanthin and cis-Beta Carotene are needed to

Table 4.4: Parameter Estimates For Main Effects Model From Result 3

	OR & 95% CI	P-Value
(Intercept)	0.07 (0.03 , 0.15)	< 0.001
Age	1.05 (1.04 , 1.05)	< 0.001
SES	0.81 (0.73 , 0.89)	< 0.001
Sex	0.53 (0.42 , 0.66)	< 0.001
Race:O Hisp	0.82 (0.41 , 1.62)	0.570
Race:White	0.37 (0.25 , 0.53)	< 0.001
Race:Black	1.22 (0.78 , 1.9)	0.389
Race:Other	0.88 (0.52 , 1.48)	0.630
Smoking:Current Smoker	2.6 (1.81 , 3.71)	< 0.001
Smoking:Former Smoker	0.93 (0.69 , 1.24)	0.611
zl_Cryptoxanthin	0.94 (0.81 , 1.1)	0.478
zl_BCarotene	0.9 (0.76 , 1.06)	0.213

Figure 4.3: B-cryptoxanthin, cis-beta carotene, and smoking



show a decrease trajectory of predicted probability of periodontal disease with beta-carotene, while those low levels of beta-cryptoxanthin show an upward trajectory of predicted probability.

4.3.2 Results 4-5: Two risk factors/confounders & one toxins or dietary nutrient

Result 4: 3-Way Interaction: Oxychlordane, Age, and SES

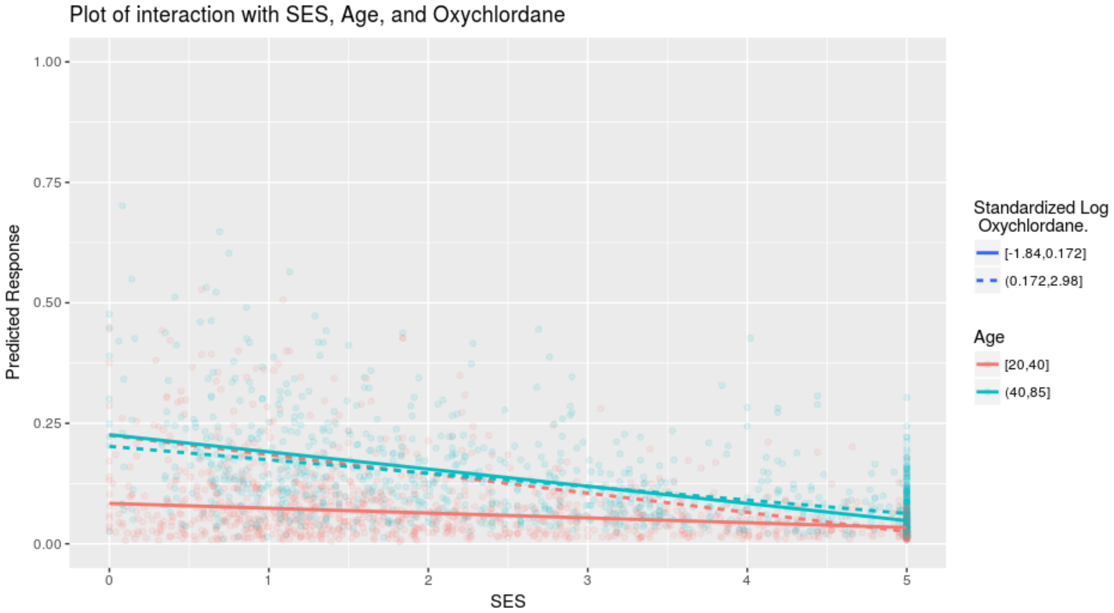
As table 4.5 shows, Oxychlordane was not statistically significant ($p < 0.05$) for the main effects model. Therefore, if main effects had only been searched for, then Oxychlordane would have not been added to the model. Both Age and SES was statistically significant ($p < 0.05$) in the main effects model. By adding the three way interaction, and its lower order terms, this result suggests the effects of Oxychlordane change as the age and SES change. Those who are younger, and have lower SES see the greatest effect of Oxyclordane on their odds of periodontal disease.

Table 4.5: Parameter Estimates For Main Effects Model From Result 4

	OR & 95% CI	P-Value
(Intercept)	0.12 (0.04 , 0.38)	0.001
Age	1.02 (1 , 1.04)	0.042
SES	0.82 (0.71 , 0.94)	0.009
SEX	0.59 (0.41 , 0.84)	0.006
Race:O Hisp	1.4 (0.59 , 3.28)	0.449
Race:White	0.49 (0.27 , 0.86)	0.018
Race:Black	1.61 (0.95 , 2.71)	0.086
Race:Other	1.39 (0.63 , 3.07)	0.418
Smoking:Current Smoker	3.11 (1.85 , 5.24)	< 0.001
Smoking:Former Smoker	1.48 (0.94 , 2.32)	0.100
zl_Oxyclordane	1.29 (0.98 , 1.7)	0.081

As fig. 4.4 shows, subjects predicted probability of periodontal disease decreases as Socio Economic Status (SES) increases across all interaction subgroups. Those subjects that had the lowest trajectory of predicted probability of periodontitis with SES were those younger in Age, and lower in Oxychlordane. Those younger subjects with higher levels of Oxychlordane showed a similar trajectory of predicted probability of periodontitis with SES to those older subjects with either higher or lower levels of Oxychlordane.

Figure 4.4: Oxychlordane, Age, and SES



Result 5: 3-Way Interaction: Mercury Inorganic, Race (Black), and SES

While all race groups were tested in the interaction, only those subjects who identified as Black were shown to have a statistically significant interaction with SES and Mercury Inorganic.

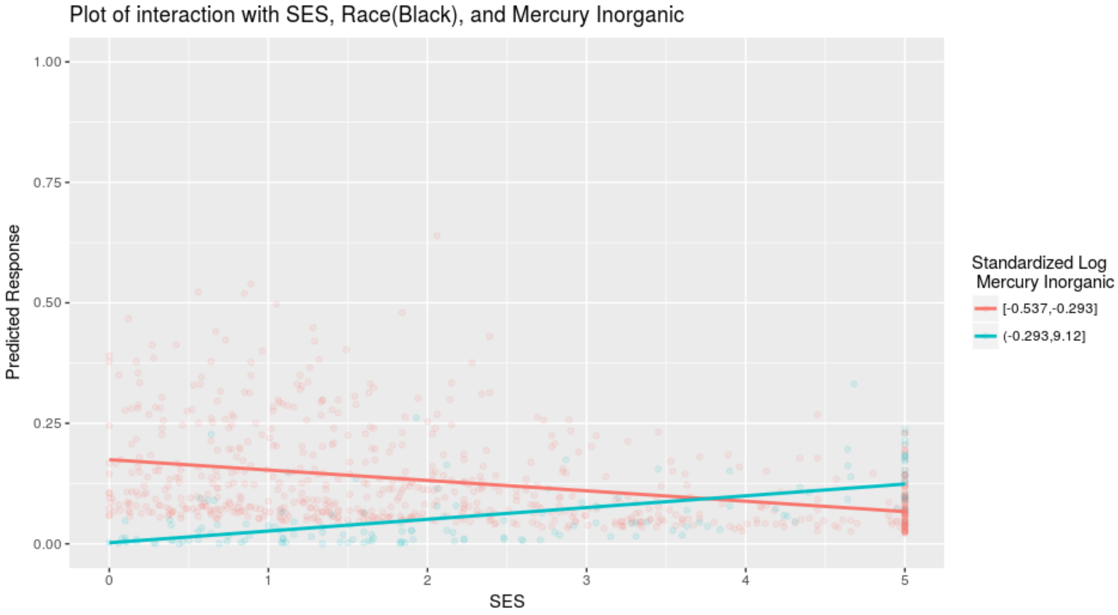
As table 4.6 shows, Mercury Inorganic, Race (Black), and SES were all statistically significant ($p < 0.05$) for the main effects model. By adding the three way interaction, and its lower order terms, this result suggests the effects of Mercury Inorganic are exacerbated by higher SES (higher Family Income to Poverty) within those who identified as Black.

Table 4.6: Parameter Estimates For Main Effects Model From Result 5

	OR & 95% CI	P-Value
(Intercept)	0.02 (0.01 , 0.05)	< 0.001
Age	1.04 (1.04 , 1.05)	< 0.001
SES	0.86 (0.78 , 0.95)	0.006
SEX	0.89 (0.6 , 1.32)	0.561
Race:O Hisp	1.33 (0.73 , 2.42)	0.358
Race:White	0.29 (0.18 , 0.45)	< 0.001
Race:Black	1.25 (0.78 , 1.99)	0.358
Race:Other	0.83 (0.45 , 1.52)	0.543
Smoking:Current Smoker	2.96 (2.03 , 4.3)	< 0.001
Smoking:Former Smoker	0.99 (0.58 , 1.7)	0.970
zl_Mercury Inorganic	0.68 (0.5 , 0.93)	0.022

As fig. 4.5 shows two different trajectories for those who identified as Black. Those subjects who had lower levels of Mercury Inorganic showed a decreasing trajectory of predicted probability of periodontitis with SES, while those with higher levels of Mercury Inorganic showed an increasing trajectory of predicted probability of periodontists with SES.

Figure 4.5: Mercury Inorganic, Race, SES



4.4 Discussion

4.4.1 Result 1: Lead, g-Tocopherol (Vit E), and Age

A positive association was found in the NHANES III dataset (1988-1994) with lead levels and prevalance of periodontal disease. Both men and women showed a positive association with periodontal disease and blood lead serum levels [49]. A Korean National Health and Nutritional Examination Survey (KNHANES) found a positive association between blood lead serum levels and periodontitis for both females and non-smokers [105]. One study examined blood lead serum levels of workers chronically exposed to lead fumes and dust in a battery plant. Those with high lead serum blood levels demonstrated increased propensity to a number of oral health issues, including periodontitis and gingivitis. These individuals also had higher levels of dental decay (carries), missing or filled teeth, and dental abrasion [**?**].

Result one indicates one major finding that has not been previously shown in other literature. As fig. 4.1 shows, as participants age, the effects of Vitamin E seem to exacerbate the effects of having higher levels of lead. The research is limited on the toxic effects of Vitamin E. However, one clinical trial examining the supplementation of Vitamin E on cardiovascular events, saw an increased hemorrhagic stroke risk in participants taking Vitamin E [127]. In addition, animal studies show that high dosage of Vitamin E can cause hemorrhage and blood coagulation [128]. Possibly, high dosage of Vitamin E supplementation is present in the survey participants. This might provide clues that Vitamin E is adversely effecting periodontal status. While this evidence is not conclusive, it does suggest a potentially interesting insight into the epidemiology of periodontitis that has yet to be explored.

4.4.2 Result 2: Alpha Carotene, Toluene, and Age

Toluene is the most commonly used as a solvent in industry. It is found in gasoline, acrylic paints, varnishes, lacquers, paint thinners, adhesives, glues, rubber cement, airplane glue, and shoe polish. A common abuse is ”glue-sniffing.” Toluene affects many biological systems including the following: central nervous system, urinary, pulmonary, cardiovascular, and gastrointestinal systems [129]. The most acute form of toxicity presents as cardiac arrest and usually occurs in individuals with an undiagnosed arrhythmia [129].

Alpha carotene is one of the caratenoids (alpha-carotene, beta-carotene,crytoxanthin, lutein, lycopene, and zeanxantin) and is considered an antioxidants. The balance between antioxidants and free radicals keeps a biological system free of inflammation. When this balance is disrupted, oxidative stress occurs and disease follows [114]. Since inflammation is implicated as the main cause of periodontal disease, researchers have focused on finding antioxidants that are protective towards periodontitis.

Interestingly, Result 2 generally shows a mitigating effect of alpha-carotene on toluene's influence on periodontitis. Also, without including the three way interaction, both Toluene and Alpha Carotene were non-significant. Previous research has failed to highlight a connection of toluene and periodontitis. The negative effect toluene has on bone metabolism might explain how higher toluene levels are associated with prevalence of periodontist. Toluene has been show to effect skeletal bones. It is possible that Toluene has a similar effect on alveolar bone. Toluene could also cause a dysregulation of the host immune response.

It is possible that alpha-carotene alleviates the build-up of IL-6, and other inflammatory markers. Individuals with low levels of alpha- and beta-carotene are more likely to have high levels of IL-6. Levels of IL-6, have been suggested as diagnostic markers for periodontal disease because of the association of these cytokines in gingival tissue of patients with periodontal disease [130].

4.4.3 Result 3: B-Cryptoxanthin, cis-Beta Carotene, and Smoking

Both B-cryptoxanthin and cis-Beta Carotene are carotenoids which act as antioxidants and fight inflammation. These carotenoids are precursors to Vitamin A and Beta-Carotene is the main source of Vitamin A in an individuals diet. Beta-Carotene is an antioxidant, which could effect could explain the protective nature of carotenoids on periodontal disease. Deficiency of Beta-Carotene and B-Cryptoxanthin has been shown to be associated with periodontitis [115]. In addition, those with low levels of

Alpha- and Beta-Carotene are more likely to have higher levels of IL-6 [131]. IL-6 is one inflammatory marker that has been suggested as a diagnostic measurement for periodontal disease. High IL-6 levels has been shown to be associated with periodontal disease [130]. The inverse association between carotenoids and inflammatory markers and periodontal disease, seems to suggest that these carotenoids mitigate the inflammation occurring in periodontal disease.

Besides a possible mitigating effect of Vitamin A precursors (carotenoids) on inflammation; Vitamin A has an anabolic effect, building of complex substances from simpler ones, on bones [132] [133]. Cryptoxanthin is also protective against osteoporosis in women. One article did a review on the mechanism by which Cryptoxanthin maintains bone homeostasis. The balance between bone formation and bone resorption is critical. B-cryptoxanthin stimulates bone formation and also inhibits bone resorption. This helps increase bone mass. According to this review, B-cryptoxanthin demonstrates "a preventative effect on bone loss in animal models for osteoporosis and in health human or postmenopausal women [134]." The maintenance of bone health via B-cryptoxanthin could also be protective in periodontal disease via the same mechanism. With these insights, result 3 seems to suggest that both B-Cryptoxanthin and Beta-Carotene support bone health in current smokers.

4.4.4 Result 4: Oxychlordane, Age, and SES

A pesticide containing a mixture of chlordanes was used beginning in 1948 until the 1980s, when its use became restricted to termite control. At the time, evidence was emerging about its toxicity [135]. In addition, information about its' half life of 10-20 years after application for termites led to its eventual removal for all uses [136]. These substances are considered persistent organic pollutants (POPs). POPs accumulate in the Arctic and then are absorbed by phytoplankton, which are eventually eaten by fish. One major concern is the ingestion of marine fish by humans. Oxychlordane, a major metabolite of chlordanes, was present in human breast-milk in similar values as arctic mammals, indicating its bio-accumulation in humans [137]. Chlordane is absorbed rapidly through the skin and eyes. Acute exposure mainly effects the neurosystem via the Central Nervous System. Other biological systems affected include: Gastrointestinal, dermal, ocular, respiratory, musculoskelatal, and hepatic [138]. Permanent alterations to the nervous system has been shown in subjects with continuous chlordane exposure [138].

Result 4 demonstrates that SES tended to decrease the predicted probability of periodontal disease across all subjects. This is a well-established association in the periodontal literature. Researchers have previously examined a variety of POPs using the NHANES 1999-2002 dataset and its relationship to periodontal disease. A positive association was found between organochlorines and periodontal disease. Young individuals with high levels of oxychlordane showed similar trajectories of predicted periodontal disease as older individuals. The predicted probability of periodontal disease with SES was shifted to reflect an older individuals if young individuals had high serum levels of oxychlordane. Considering that the previous researchers found the same trend and that periodontal disease is largely a dysregulation of an inflammatory response, it's not surprising that oxychlordane has a negative effect. These same researchers, Lee et al, found organochlorines to be the POP most strongly associated with Type II Diabetes, Insulin Resistence, and Metabolic Syndrome (all chronic diseases). [139] [140] [141].

4.4.5 Result 5: Mercury Inorganic, Race, SES

As with any heavy metals, mercury poisoning is most devastating to the central nervous system. It can lead to psychiatric disturbances, ataxia, visual loss, hearing loss, and neuropathy [142]. The other system affected most acutely by mercury poisoning is the renal system. In the United States the primary method of exposure to mercury is through ingestion of contaminated fish [142]. Additionally, mercury poisoning is associated with low birth weight and growth and development of children [143].

Result 5 demonstrated that non-Hispanic black participants with high levels of mercury inorganic showed a increase in their trajectory of predicted probability with SES while non-Hispanic blacks with low levels of mercury inorganic showed the opposite relationship. Because the relationship between SES and mercury levels is not completely consistent across all sub populations, it seems possible that non-Hispanic blacks that had low levels of mercury were simply the sub population that were not exposed to a heavy seafood diet, while those who had high levels of mercury contained the part of the sub population that was exposed to a heavy seafood diet as well as others who happened to be exposed to higher levels of mercury through their environment in another way.

4.5 Conclusion

In this paper, a novel statistical algorithm has been employed to explore statistically significant three-way interactions in a large complex data set. Because of this complexity, some of the 100 results are uninterpretable and could be considered spurious. Here, with the assistance of graphical tools, and clinical experts, each result that the algorithm yielded was checked manually for its potential for further exploration.

The importance of identifying interactions that contribute to periodontal disease cannot be understated. Due to the multi-faceted nature of periodontal disease, there are many complex relationships that could contribute to the disease. Many of these newly identified factors would not have been found if only main effects had been searched for using subset selection. These factors effects are only apparent when certain levels of another factor are high (or low) enough.

Depending on a participants exposure to environmental toxins, dietary nutrients, and their various other demographics an individual may or may not be at a considerably high risk for periodontal disease. Like many complex diseases, the risk factors are not simply additive in nature, rather, they all contribute differently depending on the presence or absence of other factors. These varying effects in light of the exposure of another factor are exactly what statistical interactions seek adjust for. By identifying consistent effect modifiers, clinicians and researchers alike will be able to better diagnose, prevent, and understand the epidemiology and etiology of periodontal disease.

Of the 33 results, five were selected on the basis of their strength in association, graphical interpretability, and their potential to yield meaningful insight into periodontal epidemiology. All five results provide interesting new insights into the relationships between environmental toxins, dietary nutrients, and demographic subgroups. These results further support the notion that periodontal disease is multifaceted in nature, and illuminate the importance of exploring higher order interactions to understand complex diseases.

In general, all five findings were compatible with the current literature. Carotenoids were protective in all the results that included them. Higher levels of pollutants, toxins, and heavy metals such as lead, mercury, and toluene were associated with an increase in odds of periodontitis. Subgroups of Race, Smoking Status, Age, and SES were also shown to have increased odds of periodontitis. Higher levels of vitamin E seemed to exacerbate the toxicity of blood lead on periodontal disease status.

Chapter 5 Conclusion

> *"The failure to identify interactive effects in regression models could lead to significant bias, misinterpretation of the results, and in some instances to incorrect public health interventions with potential adverse implications."*[122]

5.1 Statistical Interactions: Barriers

While many barriers exist to interpret interactions and understand their meaningfulness, ignoring them could lead to misinterpretation and masked effects. Most public health concerns and primary prevention, in general, involves complex, and multifaceted problems. Interactions are a potential strategy for quantifying complexity, and may illuminate subgroups of patients that are at a particularly high (or low) risk for a disease.

While interactions may be stated prior to and subsequently tested in final statistical models, testing two or three way interactions is not standard practice. Moreover, investigation of interactions tends to be limited to categorical variables. Interpretation and graphing continuous variable interactions, however, usually limits the investigator in what interactions are considered or checked. While subgroups are often of interest, they are generally explored in stratified analyses when little is known about the inter-relationships between potential risk factors, and the disease. In fact, three(or more)-way interactions are rarely explored, because these higher order interactions are even more difficult to understand and interpret then two-way interactions. Also because of the resulting number of potential combinations, the consensus avoids investigating higher-order interactions.

While these barriers, collectively, can be overwhelming, interactions may hold the greatest insights. For this reason, investigators have turned to Classification and Regression Tree's (CART). CART allows for easy intuitive interpretation because of of the visualization it provides via a branched tree diagrams. Examples of the results that CART analysis provides can be found in chapter 2.

5.2 Limitations in Diagramming Interactions

While the diagram that CART provides can be great for interpretation, its statistical downsides are sufficient to limit its potential. CART is limited in regards to how it handles the variance in the data. For instance, the statistical test used to split the variables is problematic because the standard errors it uses are estimated from only the values that are considered at that level. If there are small sample sizes, or if the sampling method is complex the standard errors are likely to be incorrect and lead to spurious splits. There are no final tests for CART, and comparing CARTs can be difficult.

Tuning parameters exist for every CART. These parameters include: lowest terminal bin size, criterion for stopping early, split parameters, and others depending on how the data are structured. CART can be very sensitive to these chosen parameters. Results can be very different from one CART to the next, even with a small change to just one of the parameters or data. These CART parameters are usually decided via simulation.

If interactions are identified in CART, it is not guaranteed that those interactions would be significant in a statistical regression model. The identification of model based interactions that are statistically significant are preferred, usually found by either exhaustively searching for them or by checking certain interactions that an investigator believes to exist.

5.3 The Feasible Solution Algorithm

5.3.1 FSA for Exploratory Analysis

The Feasible Solution Algorithm (FSA) seeks to explore interactions in datasets where little is known about the associations between covariates and the outcome of interest. FSA, unlike CART, uses statistical models as its framework for finding solutions.

The solutions that FSA produces are statistically optimal, in the sense that no one exchange to any variables under consideration will improve the chosen criterion function. These exploratory solutions can then be put through another rigorous filter (graphical overseer, opinion of a clinical expert) for their merit and/or value. As the idiom says, "a picture is worth a thousand words", a graphical view of the complex finding can go a long way in understanding the result. Using figures presented in Chapter 4, visualization tools can still be used to assist in understanding term effects in complex statistical models. This second filter, whether it be by graphical or clinical perspective, assists in translating the science for practical use.

5.3.2 A Model Based Algorithm for Identifying Interactions

Statistical models exist to correctly account for the experimental and/or survey design process which generated the data. While exploratory data analysis is useful, ultimately investigators seek a model-based approach. The FSA, uses statistical models to identify interactions, thus strengthening the toolkit available to data analysts.

5.3.3 Statistical software for Identifying Interactions

rFSA, gives users an easy to use tool to explore their data for higher order interactions or best subsets. This tool is readily available via CRAN or Github and ample documentation is provided to assist the user. Along side this package, a Shiny application has been made to allow users who are not familiar with R to upload their own data and explore best subsets and interactions. A graphical gadget has also been created in R to plot higher order interactions in statistical models. This gadget plots similar to those graphs in Chapter 4. This package, app, and gadget are regularly updated, improved upon, and are subject to change moving forward. The appendix of this dissertation highlights the codes used for the current rFSA package, version 0.1.0. All of these tools were created to assist data analyst in building better statistical models and generate new hypothesis related to disease and outcomes of interest.

5.4 A Team Science Approach

Team science is the collaborative effort of many experts from multiple disciplines to uncover new insights and ultimately advance science within a specific content domain. The research described in Chapters 2 and 4, relies on the expertise of clinical periodontists, periodontal epidemiologists, and statisticians to identify new subgroups who are particularly high (or low) risk for periodontal disease. Clinical experts were able to provide domain specific knowledge of how to define the disease and which solutions would be considered useful from a clinical perspective. Periodontal epidemiology provided insight into what toxins and/or dietary nutrients in the NHANES data should be used in the analysis, as well as how the results translate to the populations of interest. Statisticians provided insight into how to best model the data, and what methods exist to identify subgroups who are at particularly high (or low) risk. Statistical and graphical tools were able to assist the statistician, clinicians, and epidemiologists in identifying interesting solutions, and what they mean. This unity of content knowledge, data management, computer science, and statistics is data science. By taking a data and team science approach, the research communities will move closer towards personalized and precision medicine.

5.5 Insights into the Epidemiology of Periodontal Disease

The future of research in periodontics will seek to identify complex relationships of risk factors, newly discoverd biomarkers, and more precise preventative measures. By way of the FSA, this disseration work utilizes exisiting datasets to uncover interdependent relationships of known risk factors, nutrients, and environmental toxins with periodontal disease. Recent discoveries have shown that the various sub-types of periodontal disease have different etiologies, bacterial pathogens, and immunological responses [84]. Similarly, Chapter 4 highlights three way interactions between environmental toxins, vitamins, nutrients, and known risk factors which contribute to the onset of periodontal disease.

By targeting 3-way interactions with known risk factors, this dissertation provided enhanced understanding of the combination of these variables. Results 2, and 4 in Chapter 4 supports previous findings that carotenoids stimulate processes consistent with maintenance of bone homeostasis. Moreover, this result suggest that these carotenoid benefits are only present in specific sub-populations that have specific combinations of other carotenoids with the absence of smoking, and toluene. Also, in Results 3 and 4, the effects of Socio Economic Status, and Race were found to be modified by the presence or absence of toxins such as mercury and oxychlordane. While an age effect on periodontal disease has been specifically identified as cumulative in nature, Results 1-3 suggest that aging effects may be further cumulative in the presence of lead, vitamin E, oxychlordane, and SES.

5.5.1 Insights into Public health and Clinical Prevention

By identifying three way interactions related to periodontal disease, patients subgroups with specifically higher risk of periodontal disease were identified. These newly identified subgroups allow clinicians and periodontal researchers a strategy for targeting new treatments, drugs, and prevention approaches to combat the disease.

Specifically, this research suggests the added benefit of carotenoids for current smokers. One clinical approach would be to develop a periodontal supplement with a variety of carotenoids shown to assist in periodontal disease. Also, many fruits and vegetables have high levels of carotenoids and could be suggested as a type of ”prophylactic preventive” for periodontal disease patients. A diet high in fruits and vegetables has been targeted as an prevention strategy for many diseases [144].

No smoking history, and limited exposure to environmental toxins such as oxychlordane, toluene, mercury, and lead all were shown to assist in lowering the odds of periodontal disease. These effects were typically modified by the presence of carotenoids, which further reiterates the ability of carotenoids to enhance bone health.

5.6 Final Thoughts And Future Directions

Statisticians often lack the tools to adequately look for interactions in large datasets. This lack of tools limits the questions that investigators are able to ask and the discoveries that can be uncovered. The rFSA package and accompanying shiny app seek to serve as a tool available to statisticians or data scientists to identify potential interactions.

While FSA is a good first step to identify interactions, more work needs to done in exploring how to plot interactions where there are many main effects and lower order terms. Interpreting models with many variables is challenging and FSA's success hinges on the ability to adequately interpret these models.

Plans to continue to add functionality to FSA are currently underway. I plan to add more criterion functions, supported modeling types, and software. Python, JMP, and SAS are currently on my list of statistical software to support in the future. The Shiny app is also an important tool for the algorithms success. I plan to add the app to the R package, so users can use the application on their own computer where data is secure. I plan to further develop the visualization module on the app, so users can better view their results. One major benefit of FSA is its model based approach. Any statistical method can be used along with FSA to identify subsets or interactions of interest. More methods will be added in the future versions of the package. The app and R package serve as tools where researchers can explore their data and uncover new associations that could not have investigated otherwise.

Appendix: rFSA R Package

Main FSA Functions

`lmFSA`	*rFSA: Feasible Solution Algorithm (FSA) for Linear Models*

Description

A function using a Feasible Solution Algorithm to find a set of feasible solutions for a linear model of a specific form that could include mth-order interactions (Note that these solutions are optimal in the sense that no one swap to any of the variables will increase the criterion function.)

Usage

```
lmFSA(formula, data, fixvar = NULL, quad = FALSE, m = 2, numrs = 1,
cores = 1, interactions = TRUE, criterion = r.squared, minmax = "max",
...)
```

Arguments

`formula`	an object of class ”formula” (or one that can be coerced to that class): a symbolic description of the model to be fitted. See help(lm) for details.
`data`	a data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model.
`fixvar`	a variable to fix in the model. Usually a covariate that should always be included (Example: Age, Sex). Will still consider it with interactions. Default is NULL.

`quad` to include quadratic terms or not.

`m` order of terms to potentially include. If interactions is set to TRUE then m is the order of interactions to be considered. Defaults to 2. For Subset selection (interaction=F), m is the size of the subset to examine. Default is 2.

`numrs` number of random starts to perform.

`cores` number of cores to use while running. Note: Windows can only use 1 core. See mclapply for details. If function detects a Windows user it will automatically set cores=1.

`interactions` T or F for whether to include interactions in model. Defaults to FALSE.

`criterion` which criterion function to either maximize or minimize. For linear models one can use: r.squared, adj.r.squared, cv5.lmFSA (5 Fold Cross Validation error), cv10.lmFSA (10 Fold Cross Validation error), apress (Allen's Press Statistic), int.p.val (Interaction P-value), AIC, BIC.

`minmax` whether to minimize or maximize the criterion function

`...` arguments to be passed to the lm function

Details

PLEASE NOTE: make sure categorical variables are factors or characters otherwise answers will not reflect the variable being treated as a continuous variable.

Value

returns a list of solutions and table of unique solutions. $solutions is a matrix of fixed terms, start position, feasible solution, criterion function value (p-value of interaction), and number of swaps to solution. $table is a matrix of the unique feasible solutions and how many times they occured out of the number of random starts chosen. It also returns any warning messages with these solutions in the last column. $efficiency is text comparing how many models you ran during your FSA search compared to how many you would have done with exhaustive search. Note: The FSA algorithm takes additional time to run on top of the model checks that were done during the algorithm. This additional time is approximately 15

Examples

```
#use mtcars package see help(mtcars)
data(mtcars)
colnames(mtcars)
fit<-lmFSA(formula="mpg~hp+wt",data=mtcars,fixvar="hp",
quad=FALSE,m=2,numrs=10,cores=1)
print(fit) #print formulas of fitted models
summary(fit) #review
```

Code

```
lmFSA = function(formula,data,fixvar = NULL,quad = FALSE,m = 2,numrs = 1,
cores = 1,interactions = TRUE,criterion = r.squared,minmax = "max",...) {
if(identical(criterion,bdist)){return(show("Sorry the criterion function you listed cannot be used with lmFSA."))}
formula <- as.formula(formula)
fit <- lm(formula,data = data,...)
yname <- all.vars(formula)
if (!all(c(yname,fixvar) %in% colnames(data))) {
return(
show(
"Sorry, one of the variables you specified in your formula or fixvar is not a name for a column in the data you specified. Please try again."
)
)
}

originalnames <- colnames(data)
data <- data.frame(data)
lhsvar <- yname[1]

if (.Platform\$OS.type == "unix") {
} else {
```

```
cores = 1
}


ypos <- which(colnames(data) == lhsvar)
startvar <- NULL
xdata <- data[,-ypos]
ydata <- data[,ypos]
newdata <- data.frame(cbind(ydata,xdata))
fixpos <- which(colnames(xdata) %in% fixvar)
if (length(fixpos) == 0) {
fixpos = NULL
}

history <- matrix(rep(NA,numrs * (2 * m + 3)),ncol = ((2 * m + 3)))
history[,1:m] <- rstart(m = m,nvars = (dim(newdata)[2] - 1),numrs = numrs)
curpos <- which(colnames(xdata) %in% startvar[-1])
if (length(curpos) != 0) {
history <- rbind(c(curpos,rep(NA,length(curpos) + 2)),history)
}

fsa <- function(i,history,...) {
cur <- history[i,1:m]
last <- rep(NA,m)
numswap <- 0
memswap <- NULL
if (minmax == "max") {
last.criterion <- (-Inf)
}
if (minmax == "min") {
last.criterion <- (Inf)
}
checks <- 0
while (!identical(cur,last) && !identical(c(cur[2],cur[1]),last)) {
last <- cur
if (numswap == 0) {
moves <- swaps(cur = cur,n = dim(xdata)[2],quad = quad)
}
if (numswap > 0) {
moves <-
nextswap(
curpos = cur,n = dim(xdata)[2],quad = quad,prevpos = memswap
)\$nswaps
}
if (dim(moves)[2] == 0) {
```

```r
moves <- t(t(last))
}
if (interactions == T) {
form <-
function(j)
formula(paste0(
colnames(newdata)[1],"~",paste0(fixvar,sep = "+"),paste(colnames(xdata)[moves[,j]],collapse = "*")
),sep = "")
}
if (interactions == F) {
form <-
function(j)
formula(paste0(
colnames(newdata)[1],"~",paste0(fixvar,sep = "+"),paste(colnames(xdata)[moves[,j]],collapse = "+")
),sep = "")
}
tmp <-
parallel::mclapply(
X = 1:dim(moves)[2],FUN = function(k)
criterion(lm(form(k),data = newdata,...)),mc.cores = cores
)
checks <- checks + dim(moves)[2]
if (minmax == "max") {
cur <- moves[,which.max.na(unlist(tmp))[1]]
cur.criterion <- unlist(tmp[which.max.na(unlist(tmp))[1]])
if (last.criterion > cur.criterion) {
cur <- last.pos
cur.criterion <- last.criterion
}
}
if (minmax == "min") {
cur <- moves[,which.min.na(unlist(tmp))[1]]
cur.criterion <- unlist(tmp[which.min.na(unlist(tmp))[1]])
if (last.criterion < cur.criterion) {
cur <- last.pos
cur.criterion <- last.criterion
}
}
numswap <- numswap + 1
last1 <- last
last.criterion <- cur.criterion
last.pos <- cur
memswap <- unique(c(memswap,last1))
```

```
}
history[i,(1 + m):(2 * m)] <- cur
history[i,(dim(history)[2] - 2)] <- cur.criterion
history[i,(dim(history)[2] - 1)] <- numswap - 1
history[i,(dim(history)[2])] <- checks
return(history[i,])
}
solutions <-
matrix(unlist(lapply(
1:numrs,FUN = function(i)
fsa(i,history)
)),ncol = dim(history)[2],byrow = T)
solutions[,1:(2 * m)] <-
matrix(colnames(newdata)[c(solutions[,1:(2 * m)] + 1)],ncol = (2 * m))
solutions <- data.frame(solutions)
colnames(solutions) <-
c(
paste("start",1:m,sep = "."),paste("best",1:m,sep = "."),"criterion","swaps",
    "checks"
)
solutions\$criterion <-
as.numeric(levels(solutions\$criterion))[solutions\$criterion]
solutions\$swaps <-
as.numeric(levels(solutions\$swaps))[solutions\$swaps]
solutions\$checks <-
as.numeric(levels(solutions\$checks))[solutions\$checks]
if (length(fixvar) != 0) {
solutions <-
data.frame(fixvar = matrix(
rep(x = fixvar,dim(solutions)[1]),nrow = dim(solutions)[1],byrow = T
),solutions)
}
solutions <- solutions
a <- solutions[,(length(fixvar) + m + 1):(length(fixvar) + m + 1 + m)]
b <- unique(t(apply(a,sort,MARGIN = 1)),MARGIN = 1)
a <- t(apply(a,sort,MARGIN = 1))
c <- cbind(b,0)
for (i in 1:dim(b)[1]) {
for (j in 1:dim(a)[1]) {
c[i,(m + 2)] <-
sum(as.numeric(c[i,(m + 2)]) + as.numeric(identical(a[j,],b[i,])))
}
}
tableres <- data.frame(cbind(c),stringsAsFactors = F)
colnames(tableres)[(dim(tableres)[2])] <- "times"
colnames(tableres)[2:(dim(tableres)[2] - 1)] <-
paste("Var",1:m,sep = "")
colnames(tableres)[1] <- "criterion"

call <- mget(names(formals()),sys.frame(sys.nframe()))
ls <-
list(
originalfit = fit,call = call,solutions = solutions,table = tableres,
    efficiency =
paste(
"You did:",sum(solutions\$checks)," model checks compared to ",choose(n = dim
    (xdata)[2],k = m)," checks you would have done with exhaustive search."
)
)
class(ls) <- "FSA"
invisible(ls)
return(ls)
}
```

`glmFSA` *rFSA: Feasible Solution Algorithm (FSA) for Generalized Linear Models*

Description

A function using a Feasible Solution Algorithm to find a set of feasible solutions for a generalized linear model of a specific form that could include mth-order interactions (Note that these solutions are optimal in the sense that no one swap to any of the variables will increase the criterion function.)

Usage

```
glmFSA(formula, data, fixvar = NULL, quad = FALSE, m = 2, numrs = 1,
cores = 1, interactions = TRUE, criterion = AIC, minmax = "min",
family = "binomial", ...)
```

Arguments

`formula` an object of class ”formula” (or one that can be coerced to that class): a symbolic description of the model to be fitted. See help(glm) for details.

`data` a data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model.

`fixvar` a variable to fix in the model. Usually a co-variate that should always be included (Example: Age, Sex). Will still consider it with interactions. Default is NULL.

`quad` to include quadratic terms or not.

`m` order of terms to potentially include. If interactions is set to TRUE then m is the order of interactions to be considered. Defaults to 2. For Subset selection (interaction=F), m is the size of the subset to examine. Default is 2.

`numrs` number of random starts to perform.

`cores` number of cores to use while running. Note: Windows can only use 1 core. See mclapply for details. If function detects a Windows user it will automatically set cores=1.

`interactions`

T or F for whether to include interactions in model. Defaults to FALSE.

`criterion` which criterion function to either maximize or minimize. For linear models one can use: apress (Allens Press Statistic), int.p.val (Interaction p-value), AIC, BIC.

`minmax` whether to minimize or maximize the criterion function

`family` family argument passed to glm. A description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function.

... arguments to be passed to the glm function

Details

PLEASE NOTE: make sure categorical variables are factors or characters otherwise answers will not reflect the variable being treated as a continuous variable.

Value

returns a list of solutions and table of unique solutions. $solutions is a matrix of fixed terms, start position, feasible solution, criterion function value (p-value of interaction), and number of swaps to solution. $table is a matrix of the unique feasible solutions and how many times they occurred out of the number of random starts chosen. It also returns any warning messages with these solutions in the last column. $efficiency is text comparing how many models you ran during your FSA search compared to how many you would have done with exhaustive search. Note: The FSA algorithm takes additional time to run on top of the model checks that were done during the algorithm. This additional time is approximately 15

Examples

```
dat<-read.csv("http://tinyurl.com/zq7l775",header = FALSE)
colnames(dat)<-c("Class","Age","Sex","Sterioid","Antivirals",
"Fatigue","Malaise","Anorexia","Liver Big",
"Liver Firm","Spleen Palpable","Spiders",
"Ascites","Varices","Bilirubin","Alk Phosphate",
"Sgot","Albumin","Protime","Histology")
dat<-as.matrix(dat)
dat[which(dat=="?")]=NA
dat<-data.frame(dat)
dat[,c(2,15,16,17,18,19)]<-lapply(X = dat[,c(2,15,16,17,18,19)],
as.numeric)
colnames(dat)
fit<-glmFSA(formula="Class~Age+Sgot*Albumin",data=dat,fixvar="Age",
quad=FALSE,m=2,
numrs=10,family="binomial",cores=1)
```

Code

```
glmFSA = function(formula,data,fixvar = NULL,quad = FALSE,m = 2,numrs = 1, cores = 1,
interactions = TRUE,criterion = AIC,minmax = "min",family =
"binomial",...) {
if(identical(criterion,r.squared)|identical(criterion,r.squared)){return(show("Sorry the criterion function you listed cannot be used with glmFSA."))}
formula <- as.formula(formula)
fit <- glm(formula,data = data,family = family,...)
yname <- all.vars(formula)
if (!all(c(yname,fixvar) %in% colnames(data))) {
return(
show(
"Sorry, one of the variables you specified in your formula or fixvar is not a name for a column in the data you specified. Please try again."
)
)
}
originalnames <- colnames(data)
data <- data.frame(data)
lhsvar <- yname[1]

if (.Platform\$OS.type == "unix") {
} else {
cores = 1
}


ypos <- which(colnames(data) == lhsvar)
startvar <- NULL
xdata <- data[,-ypos]
ydata <- data[,ypos]
newdata <- data.frame(cbind(ydata,xdata))
fixpos <- which(colnames(xdata) %in% fixvar)
if (length(fixpos) == 0) {
fixpos = NULL
```

```
    }

    history <- matrix(rep(NA,numrs * (2 * m + 3)),ncol = ((2 * m + 3)))
    history[,1:m] <- rstart(m = m,nvars = (dim(newdata)[2] - 1),numrs = numrs)
    curpos <- which(colnames(xdata) %in% startvar[-1])
    if (length(curpos) != 0) {
    history <- rbind(c(curpos,rep(NA,length(curpos) + 2)),history)
    }

    fsa <- function(i,history,...) {
    cur <- history[i,1:m]
    last <- rep(NA,m)
    numswap <- 0
    memswap <- NULL
    if (minmax == "max") {
    last.criterion <- (-Inf)
    }
    if (minmax == "min") {
    last.criterion <- (Inf)
    }
    checks <- 0
    while (!identical(cur,last) && !identical(c(cur[2],cur[1]),last)) {
    last <- cur
    if (numswap == 0) {
    moves <- swaps(cur = cur,n = dim(xdata)[2],quad = quad)
    }
    if (numswap > 0) {
    moves <-
    nextswap(
    curpos = cur,n = dim(xdata)[2],quad = quad,prevpos = memswap
    )\$nswaps
    }
    if (dim(moves)[2] == 0) {
    moves <- t(t(last))
    }
    if (interactions == T) {
    form <-
    function(j)
    formula(paste0(
    colnames(newdata)[1],"~",paste0(fixvar,sep = "+"),paste(colnames(xdata)[moves[,j]],collapse = "*")
    ),sep = "")
    }
    if (interactions == F) {
    form <-
```

```
function(j)
formula(paste0(
colnames(newdata)[1],"~",paste0(fixvar,sep = "+"),paste(colnames(xdata)[moves[,j]],collapse = "+")
),sep = "")
}
tmp <-
parallel::mclapply(
X = 1:dim(moves)[2],FUN = function(k)
criterion(glm(
form(k),data = newdata,family = family,...
)),mc.cores = cores
)
checks <- checks + dim(moves)[2]
if (minmax == "max") {
cur <- moves[,which.max.na(unlist(tmp))[1]]
cur.criterion <- unlist(tmp[which.max.na(unlist(tmp))[1]])
if (last.criterion > cur.criterion) {
cur <- last.pos
cur.criterion <- last.criterion
}
}
if (minmax == "min") {
cur <- moves[,which.min.na(unlist(tmp))[1]]
cur.criterion <- unlist(tmp[which.min.na(unlist(tmp))[1]])
if (last.criterion < cur.criterion) {
cur <- last.pos
cur.criterion <- last.criterion
}
}
numswap <- numswap + 1
last1 <- last
last.criterion <- cur.criterion
last.pos <- cur
memswap <- unique(c(memswap,last1))
}
history[i,(1 + m):(2 * m)] <- cur
history[i,(dim(history)[2] - 2)] <- cur.criterion
history[i,(dim(history)[2] - 1)] <- numswap - 1
history[i,(dim(history)[2])] <- checks
return(history[i,])
}
solutions <-
matrix(unlist(lapply(
1:numrs,FUN = function(i)
```

```
fsa(i,history)
)),ncol = dim(history)[2],byrow = T)
solutions[,1:(2 * m)] <-
matrix(colnames(newdata)[c(solutions[,1:(2 * m)] + 1)],ncol = (2 * m))
solutions <- data.frame(solutions)
colnames(solutions) <-
c(
paste("start",1:m,sep = "."),paste("best",1:m,sep = "."),"criterion","swaps", "checks"
)
solutions\$criterion <-
as.numeric(levels(solutions\$criterion))[solutions\$criterion]
solutions\$swaps <-
as.numeric(levels(solutions\$swaps))[solutions\$swaps]
solutions\$checks <-
as.numeric(levels(solutions\$checks))[solutions\$checks]
if (length(fixvar) != 0) {
solutions <-
data.frame(fixvar = matrix(
rep(x = fixvar,dim(solutions)[1]),nrow = dim(solutions)[1],byrow = T
),solutions)
}
solutions <- solutions
a <- solutions[,(length(fixvar) + m + 1):(length(fixvar) + m + 1 + m)]
b <- unique(t(apply(a,sort,MARGIN = 1)),MARGIN = 1)
a <- t(apply(a,sort,MARGIN = 1))
c <- cbind(b,0)
for (i in 1:dim(b)[1]) {
for (j in 1:dim(a)[1]) {
c[i,(m + 2)] <-
sum(as.numeric(c[i,(m + 2)]) + as.numeric(identical(a[j,],b[i,])))
}
}
tableres <- data.frame(cbind(c),stringsAsFactors = F)
colnames(tableres)[(dim(tableres)[2])] <- "times"
colnames(tableres)[2:(dim(tableres)[2] - 1)] <-
paste("Var",1:m,sep = "")
colnames(tableres)[1] <- "criterion"

call <- mget(names(formals()),sys.frame(sys.nframe()))
ls <-
list(
originalfit = fit,call = call,solutions = solutions,table = tableres, efficiency =
paste(
"You did:",sum(solutions\$checks)," model checks compared to ",choose(n = dim(xdata)[2],k = m)," checks you would have done with exahstive search."
)
)
class(ls) <- "FSA"
invisible(print(ls))
return(ls)

}
```

S3 Supporting Functions

`fitmodels` *Model fitting function for FSA solutions*

Description

Model fitting function for FSA solutions

Usage

```
fitmodels(object, ...)
```

Arguments

`object`	FSA object to construct models on.
`...`	other parameters passed to lm or glm. See help(lm) or help(glm) for other potential arguments

Value

list of FSA models that have been fitted.

Examples

```
#use mtcars package see help(mtcars)

data(mtcars)

colnames(mtcars)

fit<-lmFSA(formula="mpg~hp*wt",data=mtcars,fixvar="hp",

quad=FALSE,m=2,numrs=10,save\_solutions=FALSE,cores=1)

fitmodels(fit)
```

Code

```
fitmodels <- function(object,...) {
stopifnot(inherits(object, "FSA"))
resls <- list()
one<-capture.output(forms <- print(object))
if (is.null(object\$call\$fam)) {
for (i in 1:(dim(object\$table)[1] + 1)) {
resls[[i]] <- lm(forms\$Formula[[i]],data = object\$call\$data,...)
}

} else{
for (i in 1:(dim(object\$table)[1] + 1)) {
resls[[i]] <-
glm(forms\$Formula[[i]],data = object\$call\$data,family = object\$call\$fam,...)
}
}
return(resls)
}
```

fitted.FSA *Fitted Values for FSA solutions*

Description

Fitted Values for FSA solutions

Usage

```
## S3 method for class 'FSA'
fitted(object, ...)
```

Arguments

`object` FSA object to get fitted values from.

`...` other parameters passed to fitmodels or fitted function. See help(fitmodels) or help(fitted) for assistance.

Value

list of fitted values from each FSA model.

Examples

```
#use mtcars package see help(mtcars)
data(mtcars)
colnames(mtcars)
fit<-lmFSA(formula="mpg~hp*wt",data=mtcars,fixvar="hp",
quad=FALSE,m=2,numrs=10,save\_solutions=FALSE,cores=1)
fitted(fit)
```

Code

```
fitted.FSA <- function(object,...) {
stopifnot(inherits(object, "FSA"))
fm <- fitmodels(object,...)
res <- list()
for (i in 1:length(fm)) {
res[[i]] <- fitted(fm[[i]],...)
}
return(res)
}
```

`plot.FSA` *Diagnostic Plots for FSA solutions*

Description

Diagnostic Plots for FSA solutions

Usage

```
## S3 method for class 'FSA'
plot(x, ask = F, easy = T, ...)
```

Arguments

`x`	FSA object to see diagnostic plots on.
`ask`	logical; if TRUE, the user is asked before each plot. See help(plot.lm).
`easy`	logical; should diagnostic plots be presented in easy to read format?
`...`	arguments to be passed to other functions.

Value

diagnostic plots to plot window.

Examples

```
#use mtcars package see help(mtcars)
data(mtcars)
colnames(mtcars)
fit<-lmFSA(formula="mpg~hp*wt",data=mtcars,fixvar="hp",
quad=FALSE,m=2,numrs=10,save\_solutions=FALSE,cores=1)
plot(x=fit)
```

Code

```
plot.FSA <- function(x,ask = F,easy = T,...) {
stopifnot(inherits(x, "FSA"))
fm <- fitmodels(x)
if (length(fm) < 4) {
dm <- length(fm)
} else
dm <- 4
if (easy == F) {
par(mfrow = c(1,4))
}
else{
par(mfrow = c(dm,4))
}
for (i in 1:length(fm)) {
one<-capture.output(pfit<-print(x))
plot(fm[[i]],ask = ask,main = rownames(pfit)[i])
}
}
```

`predict.FSA` *Prediction function for FSA solutions*

Description

Prediction function for FSA solutions

Usage

```
## S3 method for class 'FSA'
predict(object, ...)
```

Arguments

`object` FSA object to conduct predictions on.

`...` other parameters passed to fitmodels or predict functions. See help(fitmodels) or help(predict) for assistance.

Value

list of predicted values from each FSA model.

Examples

```
#use mtcars package see help(mtcars)
data(mtcars)
colnames(mtcars)
fit<-lmFSA(formula="mpg~hp*wt",data=mtcars,fixvar="hp",
quad=FALSE,m=2,numrs=10,save\_solutions=FALSE,cores=1)
predict(fit)
predict(fit,newdata=mtcars[1:15,])
```

`predict.FSA` *Prediction function for FSA solutions*

Description

Prediction function for FSA solutions

Usage

```
## S3 method for class 'FSA'
predict(object, ...)
```

Arguments

`object`	FSA object to conduct predictions on.
`...`	other parameters passed to fitmodels or predict functions. See help(fitmodels) or help(predict) for assistance.

Value

list of predicted values from each FSA model.

Examples

```
predict.FSA <- function(object,...) {
stopifnot(inherits(object, "FSA"))
fm <- fitmodels(object,...)
res <- list()
for (i in 1:length(fm)) {
res[[i]] <- predict(fm[[i]],...)
}
return(res)
}
```

`print.FSA` *Printing function for FSA solutions*

Description

Printing function for FSA solutions

Usage

```
## S3 method for class 'FSA'
print(x, ...)
```

Arguments

`x`	FSA object to print details about.
`...`	arguments to be passed to other functions.

Value

list of Feasible Solution Formula's, Original Fitted model formula and criterion function and times converged to details.

Examples

```
#use mtcars package see help(mtcars)
data(mtcars)
colnames(mtcars)
fit<-lmFSA(formula="mpg~hp*wt",data=mtcars,fixvar="hp",
quad=FALSE,m=2,numrs=10,save\_solutions=FALSE,cores=1)
print(fit)
```

Code

```
print.FSA <- function(x,...) {
stopifnot(inherits(x, "FSA"))
vars <-
x\$table[,-which(colnames(x\$table) %in% c("criterion","times","fixvar"))]
orgvars <- all.vars(x\$call\$formula)
if (x\$call\$interactions == T) {
form <-
function(j)
paste0(orgvars[1]," ~ ",if (!is.null(x\$call\$fixvar)) {
paste0(x\$call\$fixvar,collapse = " + ")
},if (!is.null(x\$call\$fixvar)) {
" + "
},paste(vars[j,],collapse = " * "),sep = "")
}
if (x\$call\$interactions == F) {
form <-
function(j)
paste0(orgvars[1]," ~ ",if (!is.null(x\$call\$fixvar)) {
paste0(x\$call\$fixvar,collapse = " + ")
},if (!is.null(x\$call\$fixvar)) {
" + "
},paste(vars[j,],collapse = " + "),sep = "")
}
tab <- cbind(matrix(lapply(
X = 1:dim(x\$table)[1],FUN = form
)),x\$table\$criterion,x\$table\$times)
tab <- rbind(c(
Reduce(paste0,deparse(formula(x\$originalfit))),x\$call\$criterion(x\$
    originalfit),NA
),tab)
tab <- data.frame(tab)
cname <- formals(x\$call\$criterion)\$name
if (is.null(cname)) {
cname = "criterion"
}
colnames(tab) <- c("Formula", cname, "Times FS")
tab[,2] <- as.numeric(unlist(tab[,2]))
rownames(tab) <-
c("Original Fit   ",paste("FS",1:dim(x\\$table)[1],sep = ""))
tab <- data.frame(tab)
print(tab)
}
```

`summary.FSA` *Summary function for FSA solutions*

Description

Summary function for FSA solutions

Usage

```
## S3 method for class 'FSA'
summary(object, ...)
```

Arguments

`object`	FSA object to see summaries on.
`...`	arguments to be passed to other functions.

Value

list of summarized lm or glm output.

Examples

```
#use mtcars package see help(mtcars)
data(mtcars)
colnames(mtcars)
fit<-lmFSA(formula="mpg~hp*wt",data=mtcars,fixvar="hp",
quad=FALSE,m=2,numrs=10,save\_solutions=FALSE,cores=1)
summary(fit)
```

Code

```
summary.FSA <- function(object,...) {
fm <- fitmodels(object)
sumresls <- list()
for (i in 1:length(fm)) {
sumresls[[i]] <- summary(fm[[i]])
}
return(sumresls)
}
```

Criterion Functions

r.squared *An rFSA Criterion Function.*

Description

rFSA Criterion Function to compute R squared.

Usage

```
r.squared(model, name = "R Squared")
```

Arguments

`model`	lm or glm fit to be passed.
`name`	passed to print.FSA

`adj.r.squared` *An rFSA Criterion Function.*

Description

rFSA Criterion Function to compute Adjusted R-Squared.

Usage

```
adj.r.squared(model, name = "Adj R Squared")
```

Arguments

`model`	lm or glm fit to be passed.
`name`	passed to print.FSA

Code

```
adj.r.squared <- function(model,name = "Adj R Squared") {
#Adjusted R squared
if(is.null(model\$family[[1]])){return(summary(model)\$adj.r.squared)
} else return(1.1)
}
```

`rmse` *An rFSA Criterion Function.*

Description

rFSA Criterion Function to compute Root Mean Squared Error.

Usage

```
rmse(model, name = "RMSE")
```

Arguments

`model`	lm or glm fit to be passed.
`name`	passed to print.FSA

Code

```
rmse <- function(model,name = "RMSE") {
#Root Mean Squared Error
sqrt(mean(model\$residuals ^ 2))
}
```

`apress` *An rFSA Criterion Function.*

Description

rFSA Criterion Function to Allen's Press Statistic.

Usage

```
apress(model, name = "PRESS")
```

Arguments

`model` lm or glm fit to be passed.

`name` passed to print.FSA

Code

```
apress <- function(model, name = "PRESS") {
#Allen's PRESS statistic
presid <- resid(model)/(1 - influence(model)\\$hat)
sum(presid^2)
}
```

`int.p.val` *An rFSA Criterion Function.*

Description

rFSA Criterion Function to compute Liklihood Ratio Test Statistics p-value for the largest order interation term.

Usage

```
int.p.val(model, name = "Interaction P-Value")
```

Arguments

`model` lm or glm fit to be passed.

`name` passed to print.FSA

Code

```
int.p.val <- function(model,name = "Interaction P-Value") {
if((is.null(model\$call\$family)|is.null(model\$family[[1]])) & !(length(grep
    (pattern = ":",names(model\$coefficients)))==0)){return(tail(anova(model,
    test="LRT")[,5],2)[1])
}  else if((is.null(model\$call\$family)|is.null(model\$family[[1]])) & (length(grep(pattern = ":",names(model\$coefficients)))==0)){
return(0)
} else if((model\$call\$family=="binomial"|model\$family[[1]]=="binomial") & !(length(grep(pattern = ":",names(model\$coefficients)))==0)){
return(tail(anova(model,test = "LRT")[,5],1))
} else if(!(length(grep(pattern = ":",names(model\$coefficients)))==0)) {
return(max(summary(model)\$coef[grep(pattern = ":",names(model\$coefficients)),4]))
} else return(0)
}
```

bdist *An rFSA Criterion Function.*

Description

rFSA Criterion Function to compute the Bhattacharyya distance.

Usage

```
bdist(model, name = "B Distance")
```

Arguments

`model`	lm or glm fit to be passed.
`name`	passed to print.FSA

Code

```
    bdist <- function(model,name = "B Distance") {
    if(length(grep(":",names(model\$coefficients)))==0){return(0)} #no interaction in model to check
    if(is.null(model\$family)){return(0)} #not logistic regression

    nam<-c(all.vars(model\$formula)[1],strsplit(names(model\$coefficients)[grep(":",names(model\$coefficients))],"[:]")[[1]])
    tmp_dat <- eval(model\$model)
    y <- tmp_dat[,nam[1]]
    x1 <- tmp_dat[,nam[2]]
    x2 <- tmp_dat[,nam[3]]

    X <- cbind(x1,x2)
    a <- X[which(y == 0),]
    b <- X[which(y == 1),]

    mu11 <- mean(a[,1])
    mu12 <- mean(a[,2])
    mu21 <- mean(b[,1])
    mu22 <- mean(b[,2])

    var1_11 <- var(a[,1])
    var1_22 <- var(a[,2])
    var1_12 <- cov(a[,1],a[,2])

    var2_11 <- var(b[,1])
    var2_22 <- var(b[,2])
    var2_12 <- cov(b[,1],b[,2])

    mu1 <- c(mu11,mu12)
    mu2 <- c(mu21,mu22)
    sig1 <- matrix(c(var1_11,var1_12,var1_12,var1_22),ncol = 2)
    sig2 <- matrix(c(var2_11,var2_12,var2_12,var2_22),ncol = 2)

    sig <- (sig1 + sig2) / 2
    a <- sig[1,1]
    b <- sig[1,2]
    c <- sig[2,1]
    d <- sig[2,2]
    temp <- matrix(c(d,-c,-b,a),ncol = 2)
    sig.inv <- 1 / (a * d - b * c) * temp
    distance <-
    1 / 8 * (t(mu1 - mu2) %*% sig.inv %*% (mu1 - mu2)) + 1 / 2 * log(det(sig) /
    sqrt(det(sig1) * det(sig2)))
    return(c(distance))
    }
```

`which.max.na` *An rFSA Internal Function.*

Description

rFSA function to compute the maximum value from a vector with NA's.

Usage

```
which.max.na(vec)
```

Arguments

`vec`	Vector to be passed.

Code

```r
which.max.na <- function(vec) {
maxval <- max(vec,na.rm = T)
which(vec == maxval)
}
```

`which.min.na` *An rFSA Internal Function.*

Description

rFSA function to compute the minimum value from a vector with NA's.

Usage

```
which.min.na(vec)
```

Arguments

`vec`	Vector to be passed.

Code

```
which.min.na <- function(vec) {
minval <- min(vec,na.rm = T)
which(vec == minval)
}
```

`list.criterion` *List all included Criteria function for lmFSA and glmFSA.*

Description

List all included Criteria function for lmFSA and glmFSA.

Usage

```
list.criterion()
```

Value

list of functions and whether lmFSA or glmFSA work with those functions.

Examples

```
list.criterion()
```

Code

```
list.criterion<-function(){
show("Accepted Criteria Functions for lmFSA and glmFSA")
show("")
tab<-rbind(c("r.squared","lmFSA"),c("adj.r.squared","lmFSA"),c("AIC","lmFSA; glmFSA"),c("BIC","lmFSA; glmFSA"),
c("rmse","lmFSA; glmFSA"),c("apress","lmFSA; glmFSA"),c("int.p.val","lmFSA; glmFSA"),
c("bdist","glmFSA (only 2 way interactions)"))
colnames(tab)<-c(’Criterion’,"Accepted in")
tab<-data.frame(tab)
show(tab)
show("")
show("You can write your own criterion function too! Or use other criterion functions from other packages. Just follow the standard format used in int.p.val or apress as an example.")
}
```

Other Supporting Functions

rstart *An rFSA internal function*

Description

rFSA function to compute random starting spots for FSA.

Code

```
rstart = function(m,nvars,quad = FALSE,numrs = 1) {
t(replicate(numrs,sample(
1:nvars,size = m,replace = quad
)))
}
```

swaps *An rFSA internal function*

Description

rFSA function to compute swapping locations from current position based on FSA.

Code

```
swaps <- function(cur,n,quad = FALSE) {
m <- length(cur)
if (!quad) {
l <- (n - m) * m
possible <- matrix(rep(cur,l),nrow = l,byrow = T)

s <- 1:n
lapply(
1:m,FUN = function(i)
s <<- s[-which(s == cur[i])]
)
lapply(
1:m,FUN = function(j)
possible[((j - 1) * (n - m) + 1):((j - 1) * (n - m) + (n - m)),j] <<-
s
)
}
if (quad) {
l <- (n - 1) * m
possible <-  matrix(rep(cur,l),nrow = l,byrow = T)
s <- 1:n
smat <- matrix(rep(0,(n - 1) * m),ncol = m)
lapply(
1:m,FUN = function(i)
smat[,i] <<- s[-which(s == cur[i])]
)
lapply(
1:m,FUN = function(j)
possible[((j - 1) * (n - 1) + 1):((j - 1) * (n - 1) + (n - 1)),j] <<-
smat[,j]
)
}
vec <- cbind(cur,t(possible))
return(unique(apply(vec,sort,MARGIN = 2),MARGIN = 2))
}
```

`nextswaps` *An rFSA internal function*

Description

rFSA function to compute swapping locations from current position based on FSA and removing places that have already been visited.

Code

```
nextswap <- function(curpos,n,prevpos,quad) {
swps <- swaps(curpos,n,quad)
nextpos <- rep(FALSE,dim(swps)[2])
for (i in 1:dim(swps)[1]) {
nextpos <- nextpos + (swps[i,] %in% prevpos)
}

retSwps <- swps[,nextpos == (length(curpos) - 2)]
if (is.null(dim(retSwps))) {
retSwps <- t(t(retSwps))
}
return(list(nswaps = retSwps,prevpos = prevpos))
}
```

`svyglmFSA` *An FSA funciton for svyglm*

Description

A function to implement svyglm with FSA.

Code

```

svyglmFSA<-function (formula, data, fixvar = NULL, quad = FALSE, m = 2, numrs = 1, cores = 1, interactions = TRUE, criterion = AIC,  minmax = "min", family = "binomial", checkfeas = NULL,design, ...)
{
  if (identical(criterion, r.squared) | identical(criterion, r.squared)) {
    return(show("Sorry the criterion function you listed cannot be used with glmFSA."))
  }
  formula <- formula
```

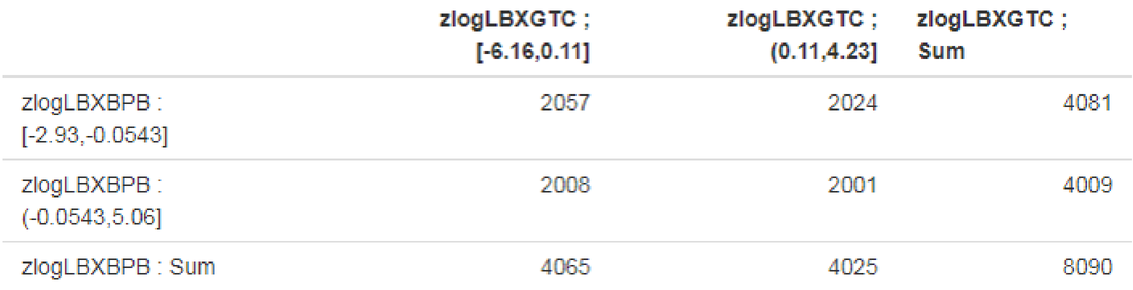
```r
fit <- svyglm(formula,design=design, family = quasibinomial() , ...)
yname <- all.vars(as.formula(formula))
if (!all(c(yname, fixvar) %in% colnames(data))) {
  return(show("Sorry, one of the variables you specified in your formula or
      fixvar is not a name for a column in the data you specified. Please try
      again."))
}
originalnames <- colnames(data)
data <- data.frame(data)
lhsvar <- yname[1]
print(lhsvar)
if (.Platform\$OS.type == "unix") {

}  else {
  cores = 1
}
ypos <- which(colnames(data) == lhsvar)
startvar <- NULL
xdata <- data[, -ypos]
ydata <- data[, ypos]
newdata <- data.frame(cbind(ydata, xdata))
fixpos <- which(colnames(xdata) %in% fixvar)
if (length(fixpos) == 0) {
  fixpos = NULL
}
history <- matrix(rep(NA, numrs * (2 * m + 3)), ncol = ((2 *
                                                           m + 3)))
if (!is.null(checkfeas)) {
  checkfeas <- which(colnames(xdata) %in% checkfeas)
  history[, 1:m] <- rbind(rstart(m = m, nvars = (dim(newdata)[2] -
                                                   1), numrs = numrs - 1), c(
                                                       checkfeas[1:m]))
}  else history[, 1:m] <- rstart(m = m, nvars = (dim(newdata)[2] -
                                                1), numrs = numrs)
curpos <- which(colnames(xdata) %in% startvar[-1])
if (length(curpos) != 0) {
  history <- rbind(c(curpos, rep(NA, length(curpos) + 2)),
                   history)
}
fsa <- function(i, history, ...) {
  cur <- history[i, 1:m]
  last <- rep(NA, m)
  numswap <- 0
  memswap <- NULL
  if (minmax == "max") {
    last.criterion <- (-Inf)
  }
  if (minmax == "min") {
    last.criterion <- (Inf)
  }
  checks <- 0
  while (!identical(cur, last) && !identical(c(cur[2],
                                                cur[1]), last)) {
    last <- cur
    if (numswap == 0) {
      print(cur);print(dim(xdata)[2]);print(quad)
      moves <- swaps(cur = cur, n = dim(xdata)[2],
                     quad = quad)
    }
    if (numswap > 0) {
      moves <- nextswap(cur= cur, n = dim(xdata)[2],quad = quad)
    }
    if (dim(moves)[2] == 0) {
      moves <- t(t(last))
    }
    if (interactions == T) {
      form <- function(j) paste0(colnames(newdata)[1],"~", paste(fixvar,
          collapse = "+"), "+", paste(colnames(xdata)[moves[,j]], collapse = "*
          "))
    }
    if (interactions == F) {
      form <- function(j) paste0(colnames(newdata)[1],"~", paste(fixvar,
          collapse = "+"), "+", paste(colnames(xdata)[moves[,j]], collapse = "+
          "))
    }
    tmp <- parallel::mclapply(X = 1:dim(moves)[2], FUN = function(k) {
      if ((sum(complete.cases(cbind(ydata, xdata[,
                                                  moves[, k]]))))/length(ydata))
                                                      > 0.05) {
        criterion(svyglm(form(k), design=design, family = quasibinomial()))
      } else {
        NA
      }
    }, mc.cores = cores)
    checks <- checks + dim(moves)[2]
    if (minmax == "max") {
      cur <- moves[, which.max.na(unlist(tmp))[1]]
      cur.criterion <- unlist(tmp[which.max.na(unlist(tmp))[1]])
      if (last.criterion > cur.criterion) {
        cur <- last.pos
        cur.criterion <- last.criterion
      }
    }
    if (minmax == "min") {
      cur <- moves[, which.min.na(unlist(tmp))[1]]
      cur.criterion <- unlist(tmp[which.min.na(unlist(tmp))[1]])
      if (last.criterion < cur.criterion) {
        cur <- last.pos
        cur.criterion <- last.criterion
      }
    }
    numswap <- numswap + 1
    last1 <- last
    last.criterion <- cur.criterion
    last.pos <- cur
    memswap <- unique(c(memswap, last1))
  }
  history[i, (1 + m):(2 * m)] <- cur
  history[i, (dim(history)[2] - 2)] <- cur.criterion
  history[i, (dim(history)[2] - 1)] <- numswap - 1
  history[i, (dim(history)[2])] <- checks
  return(history[i, ])
}
solutions <- matrix(unlist(lapply(1:numrs, FUN = function(i) fsa(i,
                                                                 history))),
                                                                     ncol = dim
                                                                     (history)
                                                                     [2], byrow
                                                                      = T)
solutions[, 1:(2 * m)] <- matrix(colnames(newdata)[c(solutions[,
                                                               1:(2 * m)] + 1)
                                                                   ], ncol = (2
                                                                    * m))
solutions <- data.frame(solutions)
colnames(solutions) <- c(paste("start", 1:m, sep = "."),
                         paste("best", 1:m, sep = "."), "criterion", "swaps",
                         "checks")
solutions\$criterion <-
as.numeric(levels(solutions\$criterion))[solutions\$criterion]
solutions\$swaps <- as.numeric(levels(solutions\$swaps))[solutions\$swaps]
solutions\$checks <- as.numeric(levels(solutions\$checks))[solutions\$checks]
if (length(fixvar) != 0) {
  solutions <- data.frame(fixvar = matrix(rep(x = fixvar,
                                              dim(solutions)[1]), nrow = dim(
                                                  solutions)[1], byrow = T),
                            solutions)
  }
  solutions <- solutions
  a <- solutions[, (length(fixvar) + m + 1):(length(fixvar) +
                                               m + 1 + m)]
  b <- unique(t(apply(a, sort, MARGIN = 1)), MARGIN = 1)
  a <- t(apply(a, sort, MARGIN = 1))
  c <- cbind(b, 0)
  for (i in 1:dim(b)[1]) {
    for (j in 1:dim(a)[1]) {
      c[i, (m + 2)] <- sum(as.numeric(c[i, (m + 2)]) +
                             as.numeric(identical(a[j, ], b[i, ])))
    }
  }
  tableres <- data.frame(cbind(c), stringsAsFactors = F)
  colnames(tableres)[(dim(tableres)[2])] <- "times"
  colnames(tableres)[2:(dim(tableres)[2] - 1)] <- paste("Var",
                                                        1:m, sep = "")
  colnames(tableres)[1] <- "criterion"
  call <- mget(names(formals()), sys.frame(sys.nframe()))
  ls <- list(originalfit = fit, call = call, solutions = solutions,
table = tableres, efficiency = paste("You did:", sum(solutions\$checks),
" model checks compared to ", choose(n = dim(xdata)[2],
k = m), " checks you would have done with exahstive search."))
class(ls) <- "FSA"
invisible(print(ls))
  return(ls)
}
```

Additional Figures and Tables

Table A1: FSA results

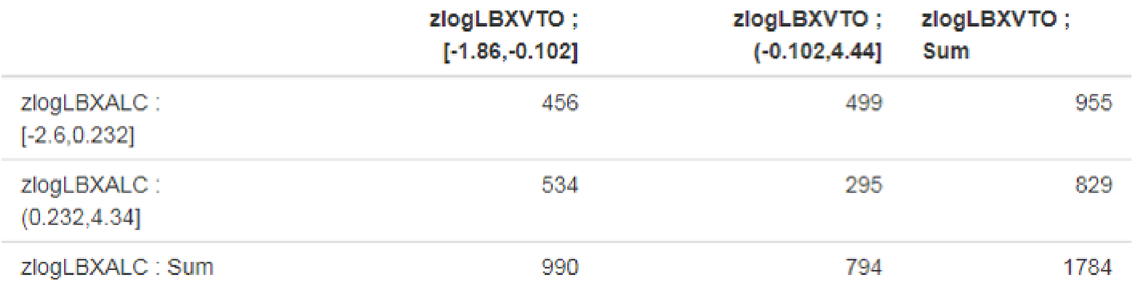
Interaction 1	Interaction 2	Interaction 3	Variable Names	P-Value	Number of Times a FS
Vitamin A	hxcb	Gender	zlLBXVIA,zlLBXHXC,RIAGENDR	< 0.0001	1
Heptachlo	DDT	Gender	zlLBXHPE,zlLBXODT,RIAGENDR	0.0050	2
Mirex	Bromoform	Gender	zlLBXMIR,zlLBXWBF,RIAGENDR	0.0027	4
hxcdd	Chlorofor	Gender	zlLBXD02,zlLBXWCF,RIAGENDR	0.0026	2
Bromodich	Styrene	Gender	zlLBXVBM,zlLBXVST,RIAGENDR	0.0003	1
Mercury I	Carbon Te	Gender	zlLBXIHG,zlLBXVCT,RIAGENDR	< 0.0001	6
Mercury	pncb	Gender	zlLBXTHG,zlLBXPCB,RIAGENDR	< 0.0001	4
Iron Seru	Tetrachlo	Age	zlLBXIRN,zlLBXV4C,RIDAGEYR	0.0070	3
Oxychlord	Benzene	Age	zlLBXOXY,zlLBXVBZ,RIDAGEYR	0.0005	5
Cotinine	Trichlore	Age	zlLBXCOT,zlLBXV3A,RIDAGEYR	0.0036	5
Lead	g-Tocophe	Age	zlLBXBPB,zlLBXGTC,RIDAGEYR	0.0085	3
Mercury	pncdf	Age	zlLBXTHG,zlLBXF03,RIDAGEYR	0.0012	3
a-caroten	Toluene	Age	zlLBXALC,zlLBXVTO,RIDAGEYR	0.0010	1
Cadmium	MTBE	SES	zlURDUCD,zlLBXWME,INDFMPIR	0.0005	1
Hexachlor	Dibromoch	SES	zlLBXHCB,zlLBXWCM,INDFMPIR	0.0013	2
o,p'-DDT	Trans-non	SES	zlLBXODT,zlLBXTNA,INDFMPIR	0.0023	5
Oxychlord	Age	SES	zlLBXOXY,INDFMPIR,RIDAGEYR	0.0005	2
pncdd	Trichloro	SES	zlLBXD01,zlLBXV3A,INDFMPIR	0.0011	1
Cadmium	Tetrachlo	SES	zlLBXBCD,zlLBXV4C,INDFMPIR	0.0012	2
hxcb	Benzene	SES	zlLBXHXC,zlLBXVBZ,INDFMPIR	0.0001	2
pncb	Xylene	SES	zlLBXPCB,zlLBXVXY,INDFMPIR	0.0076	1
Mercury I	Race	SES	zlLBXIHG,RIDRETH1,INDFMPIR	0.0021	1
a-caroten	Chlorofor	SES	zlLBXALC,zlLBXWCF,INDFMPIR	0.0001	3
Folate	Ethylbenz	Race	zlLBXRBF,zlLBXVEB,RIDRETH1	0.0013	1
Folate	o,p'-DDT	Race	zlLBXFOL,zlLBXODT,RIDRETH1	0.0004	1
g-Tocophe	Xylene	Race	zlLBXGTC,zlLBXVOX,RIDRETH1	0.0007	1
Mirex	Bromodich	Race	zlLBXMIR,zlLBXVBM,RIDRETH1	0.0052	3
p,p' DDT	Dichlorob	Race	zlLBXPDT,zlLBXVDB,RIDRETH1	0.0013	1
hpcdf	Dibromoch	Race	zlLBXF08,zlLBXWCM,RIDRETH1	0.0013	3
hxcdf	Bromodich	Race	zlLBXF05,zlLBXWBM,RIDRETH1	0.0024	1
Bromoform	MTBE	Race	zlLBXVBF,zlLBXWME,RIDRETH1	0.0004	2
Mercury I	hxcdf	Race	zlLBXIHG,zlLBXF07,RIDRETH1	0.0010	5
Vitamin E	Xylene	Race	zlLBXVIE,zlLBXVXY,RIDRETH1	0.0083	2
b-cryptox	cis-beta	Smoking	zlLBXCRY,zlLBXCBC,smoking	0.0051	3
transbeta	dibromoch	Smoking	zlLBXBEC,zlLBXWCM,smoking	0.0137	1
Dieldrin	Hexachlor	Smoking	zlLBXDIE,zlLBXHCB,smoking	0.0012	1
Mirex	Chlorofor	Smoking	zlLBXMIR,zlLBXVCF,smoking	0.0052	5
hxcdd	race	Smoking	zlLBXD04,RIDRETH1,smoking	0.0057	1
tcdd	hxcdf	Smoking	zlLBXTCD,zlLBXF07,smoking	0.0006	3
hxcdf	styrene	Smoking	zlLBXF05,zlLBXVST,smoking	0.0003	2
Carbon Te	MTBE	Smoking	zlLBXVCT,zlLBXVME,smoking	0.0031	1
Inorganic	Folate Se	Smoking	zlLBXIHG,zlLBXFOL,smoking	0.0006	3

Table A2: Lead, and g-Tocopherol (Vit E) categorical breakdown



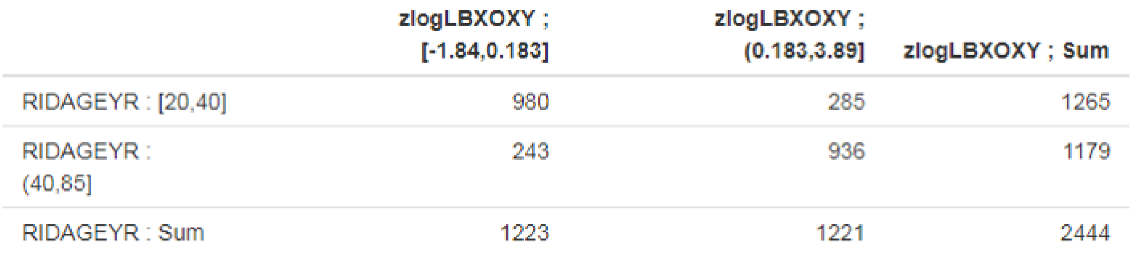
	zlogLBXGTC ; [-6.16,0.11]	zlogLBXGTC ; (0.11,4.23]	zlogLBXGTC ; Sum
zlogLBXBPB : [-2.93,-0.0543]	2057	2024	4081
zlogLBXBPB : (-0.0543,5.06]	2008	2001	4009
zlogLBXBPB : Sum	4065	4025	8090

Table A3: Alpha Carotene, and Toluene categorical breakdown



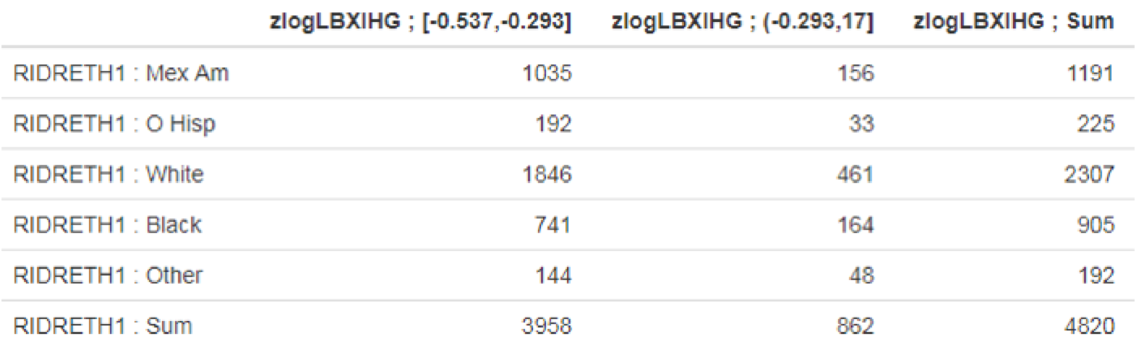
	zlogLBXVTO ; [-1.86,-0.102]	zlogLBXVTO ; (-0.102,4.44]	zlogLBXVTO ; Sum
zlogLBXALC : [-2.6,0.232]	456	499	955
zlogLBXALC : (0.232,4.34]	534	295	829
zlogLBXALC : Sum	990	794	1784

Table A4: Oxychlordane, and Age breakdown



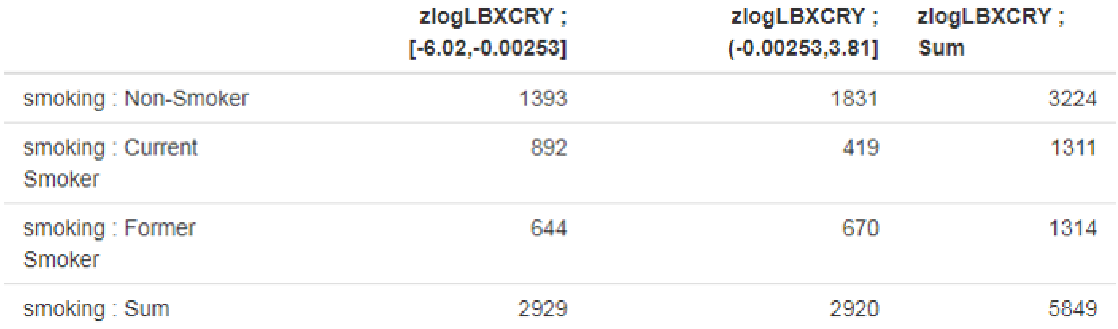
	zlogLBXOXY ; [-1.84,0.183]	zlogLBXOXY ; (0.183,3.89]	zlogLBXOXY ; Sum
RIDAGEYR : [20,40]	980	285	1265
RIDAGEYR : (40,85]	243	936	1179
RIDAGEYR : Sum	1223	1221	2444

Table A5: Race, and Mercury Inorganic categorical breakdown



	zlogLBXIHG ; [-0.537,-0.293]	zlogLBXIHG ; (-0.293,17]	zlogLBXIHG ; Sum
RIDRETH1 : Mex Am	1035	156	1191
RIDRETH1 : O Hisp	192	33	225
RIDRETH1 : White	1846	461	2307
RIDRETH1 : Black	741	164	905
RIDRETH1 : Other	144	48	192
RIDRETH1 : Sum	3958	862	4820

Table A6: Smoking, and B-Cryptoxanthin categorical breakdown



	zlogLBXCRY ; [-6.02,-0.00253]	zlogLBXCRY ; (-0.00253,3.81]	zlogLBXCRY ; Sum
smoking : Non-Smoker	1393	1831	3224
smoking : Current Smoker	892	419	1311
smoking : Former Smoker	644	670	1314
smoking : Sum	2929	2920	5849

Bibliography

[1] J Martin et al. Periodontitis severity plus risk as a tooth loss predictor. *Journal of Periodontology*, 80(2):202209, Feb 2009.

[2] Periodontal gum disease: Adults. *National Institute of Dental and Craniofacial Research*, May 2014.

[3] A I Ismail and S M Szpunar. The prevalence of total tooth loss, dental caries, and periodontal disease among Mexican Americans, Cuban Americans, and Puerto Ricans: findings from HHANES 1982-1984. *Am J Public Health*, 80 Suppl:66–70, 1990.

[4] Bidinotto et al. Change in quality of life and its association with oral health and other factors in community-dwelling elderly adults-a prospective cohort study. *Journal of the American Geriatrics Society*, 64(12):25332538, Dec 2016.

[5] E Bernab and W Marcenes. Periodontal disease and quality of life in british adults. *Journal of Clinical Periodontology*, 37(11):968972, 2010.

[6] F Hugo et al. Oral status and its association with general quality of life in older independent-living south-brazilians. *Community Dentistry and Oral Epidemiology*, 37(3):231240, Jun 2009.

[7] Periodontal disease. *Centers for Disease Control and Prevention: Periodontal Disease - Division of Oral Health*, Mar 2015.

[8] Oral health. *Oral Health Healthy People 2020.*

[9] CDC. Periodontal disease. *Centers for Disease Control and Prevention*, Oct 2015.

[10] B Pihlstrom et al. Periodontal diseases. *The Lancet*, 366(9499):18091820, 2005.

[11] P Eke et al. Update of the case definitions for populationbased surveillance of periodontitis. *Journal of Periodontology*, 83(12):14491454, 2012.

[12] P Papapanou. Periodontal diseases: epidemiology. *Annals of Periodontology*, 1996.

[13] American Academy of Periodontology. Types of gum disease. *Perio.org*.

[14] Abscess: Periodontal abscess. *ABSCESS: PERIODONTAL ABSCESS - AAP Connect*.

[15] PI Eke et al. Update on prevalence of periodontitis in adults in the united states: Nhanes 2009 to 2012. *Journal of Periodontology*, 86(5):611622, May 2015.

[16] J Albandar et al. Global epidemiology of periodontal diseases: an overview. *Periodontology 2000*, 29(1):710, 2002.

[17] JE Frencken et al. Global epidemiology of dental caries and severe periodontitis - a comprehensive review. *Journal of Clinical Periodontology*, 44:S94S105, 2017.

[18] J Albandar et al. Global epidemiology of periodontal diseases in children and young persons. *Periodontology 2000*, 29(1):153176, 2002.

[19] V John et al. Periodontal Disease and Systemic Diseases: An Update for the Clinician. *J Indiana Dent Assoc*, 95(1):16–23, 2016.

[20] V Hans et al. Delineating periodontal research: Climb is worth the view. *Journal of the International Clinical Dental Research Organization*, 8(1):14, 2016.

[21] S S Socransky, A D Haffajee, M A Cugini, C Smith, and R L Jr Kent. Microbial complexes in subgingival plaque. *J Clin Periodontol*, 25(2):134–44, 1998.

[22] R Verma et al. Porphyromonas gingivalis and Treponema denticola Mixed Microbial Infection in a Rat Model of Periodontal Disease. *Interdiscip Perspect Infect Dis*, 2010:605125, 2010.

[23] S Ji and Y Choi. Innate immune response to oral bacteria and the immune evasive characteristics of periodontal pathogens. *J Periodontal Implant Sci*, 43(1):3–11, 2013.

[24] S Ji, Y S Choi, and Y Choi. Bacterial invasion and persistence: critical events in the pathogenesis of periodontitis? *J Periodontal Res*, 50(5):570–85, 2015.

[25] N Silva et al. Host response mechanisms in periodontal diseases. *J Appl Oral Sci*, 23(3):329–55, 2015.

[26] S Steinsvoll, K Helgeland, and K Schenck. Mast cells–a role in periodontal diseases? *J Clin Periodontol*, 31(6):413–9, 2004.

[27] KS Kornman. Interleukin 1 genetics, inflammatory mechanisms, and nutrigenetic opportunities to modulate diseases of aging. *Am J Clin Nutr*, 83(2):475S–483S, 2006.

[28] G P Garlet. Destructive and protective roles of cytokines in periodontitis: a re-appraisal from host defense and tissue destruction viewpoints. *J Dent Res*, 89(12):1349–63, 2010.

[29] R J Genco. Current view of risk factors for periodontal diseases*. *Journal of Periodontology*, 67(10):10411049, 1996.

[30] P Eke et al. Risk indicators for periodontitis in us adults: Nhanes 2009 to 2012. *Journal of Periodontology*, 87(10):11741185, Jul 2016.

[31] H Shiau et al. Sex differences in destructive periodontal disease: A systematic review. *Journal of Periodontology*, 81(10):13791389, Oct 2010.

[32] G Thortan-Evans et al. Periodontitis among adults aged 30 years united states, 20092010. *CDC: Morbility and Mortality Report; Periodontitis Among Adults Aged 30 Years United States, 20092010*, Nov 2013.

[33] L Borrell et al. Socioeconomic disadvantage and periodontal disease: The dental atherosclerosis risk in communities study. *American Journal of Public Health*, 96(2):332339, Feb 2006.

[34] V Chava and R Gundala. Effect of lifestyle, education and socioeconomic status on periodontal health. *Contemporary Clinical Dentistry*, 1(1):2326, Jan 2010.

[35] T Barnea et al. Genetic polymorphisms of TNFA and IL-1A and generalized aggressive periodontitis. *Rom J Morphol Embryol*, 56(2):459–64, 2015.

[36] V Hans et al. Genetic polymorphism of Fcgamma-receptors IIa, IIIa and IIIb in South Indian patients with generalized aggressive periodontitis. *J Oral Sci*, 53(4):467–74, 2011.

[37] JM Stein et al. Human leukocyte antigen polymorphism in chronic and aggressive periodontitis among Caucasians: a meta-analysis. *J Clin Periodontol*, 35(3):183–92, 2008.

[38] X Zeng et al. Meta-Analysis of Association Between Interleukin-1beta C-511T Polymorphism and Chronic Periodontitis Susceptibility. *J Periodontol*, 86(6):812–9, 2015.

[39] K S Kornman and P J Polverini. Clinical application of genetics to guide prevention and treatment of oral diseases. *Clin Genet*, 86(1):44–9, 2014.

[40] A Johannsen et al. Smoking and inflammation: evidence for a synergistic role in chronic disease. *Periodontol 2000*, 64(1):111–26, 2014.

[41] J Haber, J Wattles, M Crowley, R Mandell, K Joshipura, and R L Kent. Evidence for cigarette smoking as a major risk factor for periodontitis. *J Periodontol*, 64(1):16–23, 1993.

[42] P Martinez-Canut, A Lorca, and R Magan. Smoking and periodontal disease severity. *J Clin Periodontol*, 22(10):743–9, 1995.

[43] C Giannopoulou et al. Effect of inflammation, smoking and stress on gingival crevicular fluid cytokine level. *J Clin Periodontol*, 30(2):145–53, 2003.

[44] S Renvert, G Dahlen, and M Wikstrom. The clinical and microbiological effects of non-surgical periodontal therapy in smokers and non-smokers. *J Clin Periodontol*, 25(2):153–7, 1998.

[45] D Souza et al. Influence of alcohol consumption on alveolar bone level associated with ligature-induced periodontitis in rats. *Braz Oral Res*, 23(3):326–32, 2009.

[46] G Ari et al. Epigenetics and periodontitis: A contemporary review. *Journal of Clinical and Diagnostic Research : JCDR*, Nov 2016.

[47] T Yang et al. Personal exposure to particulate matter and inflammation among patients with periodontal disease. *Science of The Total Environment*, 502:585589, Jan 2015.

[48] Arora et al. Association of environmental cadmium exposure with periodontal disease in u.s. adults. *Environmental Health Perspectives*, 117(5):739744, Jan 2009.

[49] M. C. P. Saraiva, R. S. Taichman, T. Braun, J. Nriagu, S. A. Eklund, and B. A. Burt. Lead exposure and periodontitis in us adults. *Journal of Periodontal Research*, 42(1):4552, 2007.

[50] D. W. Dodington, P. C. Fritz, P. J. Sullivan, and W. E. Ward. Higher intakes of fruits and vegetables, -carotene, vitamin c, -tocopherol, epa, and dha are positively associated with periodontal healing after nonsurgical periodontal therapy in nonsmokers but not in smokers. *Journal of Nutrition*, 145(11):25122519, 2015.

[51] S Najeeb et al. The role of nutrition in periodontal health: An update. *Nutrients*, 8(9):530, 2016.

[52] V John et al. Periodontal Disease and Systemic Diseases: An Update for the Clinician. *J Indiana Dent Assoc*, 95(1):16–23, 2016.

[53] S Ogbonnia. Cardiovascular risk factors: Implications in diabetes, other disease states and herbal drugs. *The Cardiovascular System - Physiology, Diagnostics and Clinical Implications*, 2012.

[54] E Lalla and P N. Papapanou. Diabetes mellitus and periodontitis: a tale of two common interrelated diseases. *Nature Reviews Endocrinology*, 7(12):738748, 2011.

[55] Atherosclerosis. *American Heart Association*, Jul 2017.

[56] R Ross. Atherosclerosis–an inflammatory disease. *N Engl J Med*, 340(2):115–26, 1999.

[57] J Danesh, R Collins, and R Peto. Chronic infections and coronary heart disease: is there a link? *Lancet*, 350(9075):430–6, 1997.

[58] M C Herzberg and M W Meyer. Effects of oral flora on platelets: possible consequences in cardiovascular disease. *J Periodontol*, 67(10 Suppl):1138–42, 1996.

[59] J Syrjanen, J Peltola, V Valtonen, M Iivanainen, M Kaste, and J K Huttunen. Dental infections in association with cerebral infarction in young and middle-aged men. *J Intern Med*, 225(3):179–84, 1989.

[60] K J Mattila, M S Nieminen, V V Valtonen, V P Rasi, Y A Kesaniemi, S L Syrjala, P S Jungell, M Isoluoma, K Hietaniemi, and M J Jokinen. Association between dental health and acute myocardial infarction. *BMJ*, 298(6676):779–81, 1989.

[61] A J Grau, F Buggle, C Ziegler, W Schwarz, J Meuser, A J Tasman, A Buhler, C Benesch, H Becher, and W Hacke. Association between acute cerebrovascular ischemia and chronic and recurrent infection. *Stroke*, 28(9):1724–9, 1997.

[62] S J Jr Arbes, G D Slade, and J D Beck. Association between extent of periodontal attachment loss and self-reported history of heart attack: an analysis of NHANES III data. *J Dent Res*, 78(12):1777–82, 1999.

[63] BG Loos. Systemic markers of inflammation in periodontitis. *J Periodontol*, 76(11 Suppl):2106–15, 2005.

[64] B Chiu. Multiple infections in carotid atherosclerotic plaques. *American Heart Journal*, 138(5):534536, 1999.

[65] W. G. Haynes. Periodontal disease and atherosclerosis: From dental to arterial plaque. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 23(8):13091311, Jan 2003.

[66] G Salvi et al. Effects of diabetes mellitus on periodontal and peri-implant conditions: update on associations and risks. *J Clin Periodontol*, 35(8 Suppl):398–409, 2008.

[67] N Chavarry et al. The relationship between diabetes mellitus and destructive periodontal disease: a meta-analysis. *Oral Health Prev Dent*, 7(2):107–27, 2009.

[68] Y Khader et al. Periodontal status of diabetics compared with nondiabetics: a meta-analysis. *J Diabetes Complications*, 20(1):59–68, 2006.

[69] B Mealey et al. Diabetes mellitus and periodontal disease. *Periodontol 2000*, 44:127–53, 2007.

[70] C Tsai et al. Glycemic control of type 2 diabetes and severe periodontal disease in the US adult population. *Community Dent Oral Epidemiol*, 30(3):182–92, 2002.

[71] L J Cianciola, B H Park, E Bruck, L Mosovich, and R J Genco. Prevalence of periodontal disease in insulin-dependent diabetes mellitus (juvenile diabetes). *J Am Dent Assoc*, 104(5):653–60, 1982.

[72] G W Taylor, B A Burt, M P Becker, R J Genco, M Shlossman, W C Knowler, and D J Pettitt. Severe periodontitis and risk for poor glycemic control in patients with non-insulin-dependent diabetes mellitus. *J Periodontol*, 67(10 Suppl):1085–93, 1996.

[73] R Demmer et al. Periodontal status and A1C change: longitudinal results from the study of health in Pomerania (SHIP). *Diabetes Care*, 33(5):1037–43, 2010.

[74] C Patel et al. An environment-wide association study (ewas) on type 2 diabetes mellitus. *PLoS ONE*, 5(5), 2010.

[75] Data mining and pattern discovery using exploratory and visualization methods for large multidimensional datasets. 2013. Theses and Dissertations–Epidemiology and Biostatistics UKy.

[76] BA Dye and G Thornton-Evans. A brief history of national surveillance efforts for periodontal disease in the United States. *J Periodontol*, 78(7 Suppl):1373–9, 2007.

[77] Catalog record: Oral health of united states adults : the national survey of oral health in u.s. employed adults and seniors, 1985-1986 : national findings. *Catalog Record: Oral health of United States adults : the... — Hathi Trust Digital Library.*

[78] A Kingman, E Morrison, H Loe, and J Smith. Systematic errors in estimating prevalence and severity of periodontal disease. *J Periodontol*, 59(11):707–13, 1988.

[79] Plan and operation of the Third National Health and Nutrition Examination Survey, 1988-94. Series 1: programs and collection procedures. *Vital Health Stat 1*, (32):1–407, 1994.

[80] L.j. Brown, J.a. Brunelle, and A. Kingman. Periodontal status in the united states, 198891: Prevalence, extent, and demographic variation. *Journal of Dental Research*, $75(2_suppl) : 672683, 1996$.

[81] J.m. Albandar, J.a. Brunelle, and A. Kingman. Destructive periodontal disease in adults 30 years of age and older in the united states, 1988-1994. *Journal of Periodontology*, 70(1):1329, 1999.

[82] B A Dye, L K Barker, R H Selwitz, B G Lewis, T Wu, C D Fryar, Y Ostchega, E D Beltran, and E Ley. Overview and quality assurance for the National Health and Nutrition Examination Survey (NHANES) oral health component, 1999-2002. *Community Dent Oral Epidemiol*, 35(2):140–51, 2007.

[83] A Dentino et al. Prevention of periodontal diseases. *Dental Clinics of North America*, 49(3):573594, 2005.

[84] J Slots. Periodontitis: facts, fallacies and the future. *Periodontology 2000*, 75(1):723, 2017.

[85] J Ebersole et al. The periodontal war: microbes and immunity. *Periodontology 2000*, 75(1):52115, 2017.

[86] A Dumitrescu. Editorial: Periodontal disease a public health problem. *Frontiers in Public Health*, 3, Aug 2016.

[87] Coverage of medicaid dental benefits for adults. *Coverage of Medicaid Dental Benefits for Adults : MACPAC*.

[88] I Lamster. Oral health care in the future: Expansion of the scope of dental practice to improve health. *Journal of Dental Education*, 81(9s), Jan 2017.

[89] B Goudey et al. High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in genome wide association studies. *Health Information Science and Systems*, 3(Suppl 1), 2015.

[90] Hawkins. The feasible solution algorithm for least trimmed squares regression. *Computational Statistics and Data Analysis*, 17:185–196, 1994.

[91] AJ Miller. Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, 147(3):389, 1984.

[92] Lumley and Miller. **leaps**: regression subset selection. 2009. R package version 2.9.

[93] Jerome F. **glmnet**:lasso and elastic-net regularized generalized linear models. 2008. R package version 2.0-5.

[94] Jiang. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, 10, 2009.

[95] Lampa. The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. *Environmental Health*, 13, 2014.

[96] Zhang and Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39, 2007.

[97] A Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Calcutta Math. Soc*, 35:99109, 1943.

[98] Vincent C. **glmulti**: Model selection and multimodel inference made easy. 2013. R package version 1.0.7.

[99] J M. Albandar. Periodontal diseases in north america. *Periodontology 2000*, 29(1):3169, 2002.

[100] D J Caplan and J A Weintraub. The oral health burden in the united states: a summary of recent epidemiologic studies. *Journal of Dental Education*, 57(12):853862, Dec 1993.

[101] Joshua Lambert. rfsa: Feasible solution algorithm for finding best subsets and interactions. *Version 0.1.0*, 2016.

[102] P Preshaw et al. Current concepts in periodontal pathogenesis. *Dental Update*, 31(10):570578, Feb 2004.

[103] Janet M. Guthmiller. *Polymicrobial diseases.* ASM Press, 2002.

[104] C P Holstege, J S Huff, A K Rowden, and R N O'Malley. Pathophysiology and etiology of lead toxicity. *Pathophysiology and Etiology of Lead Toxicity: Pharmacokinetics, Mechanisms of Toxicity, Sources of Lead Exposure*, Oct 2016.

[105] Y Won et al. Association of internal exposure of cadmium and lead with periodontal disease: a study of the fourth korean national health and nutrition examination survey. *Journal of Clinical Periodontology*, 40(2):118124, 2012.

[106] D Hicks et al. Effects of lead on growth plate chondrocyte phenotype. *Toxicology and Applied Pharmacology*, 140(1):164172, 1996.

[107] R. F. Klein. Regulation of osteoblastic gene expression by lead. *Endocrinology*, 132(6):25312537, Jan 1993.

[108] A Manocha et al. Lead as a risk factor for osteoporosis in post-menopausal women. *Indian Journal of Clinical Biochemistry*, 32(3):261265, 2016.

[109] F. E. Hemphill, M. L. Kaeberle, and W. B. Buck. Lead suppression of mouse resistance to salmonella typhimurium. *Science*, 172(3987):10311032, Apr 1971.

[110] M Mccabe et al. Lead, a major environmental pollutant, is immunomodulatory by its differential effects on cd4 t cell subsets. *Toxicology and Applied Pharmacology*, 111(1):1323, 1991.

[111] H Kishikawa, R Song, and D Lawrence. Interleukin-12 promotes enhanced resistance to infection of lead-exposed mice. *Toxicology and Applied Pharmacology*, 147(2):180189, 1997.

[112] D H Han, S Y Lim, B C Sun, S J Janket, J B Kim, D I Paik, D Paek, and H D Kim. Mercury exposure and periodontitis among a korean population: the shiwha-banwol environmental health study. *Journal of Periodontology*, 80(12):19281936, Dec 2009.

[113] Y Kim and B K Lee. Association between blood lead and mercury levels and periodontitis in the korean general population: analysis of the 2008-2009 korean national health and nutrition examination survey data. *International Archives of Occupational and Environmental Health*, 86(5):607613, Jul 2013.

[114] P. Rao et al. Free radicals and tissue damage: Role of antioxidants. *Free Radicals and Antioxidants*, 1(4):27, 2011.

[115] G Linden et al. Antioxidants and periodontitis in 60-70-year-old men. *Journal of Clinical Periodontology*, 36(10):843849, 2009.

[116] W Chen et al. On bayesian methods of exploring qualitative interactions for targeted treatment. *Statistics in Medicine*, 31(28):36933707, 2012.

[117] N Nishida et al. Determination of smoking and obesity as periodontitis risks using the classification and regression tree method. *Journal of Periodontology*, 76(6):923928, 2005.

[118] M Nunn et al. Development of prognostic indicators using classification and regression trees for survival. *Periodontology 2000*, 58(1):134142, Jan 2011.

[119] C Ramseier et al. Identification of pathogen and host-response markers correlated with periodontal disease. *Journal of Periodontology*, 80(3):436446, 2009.

[120] Huja et al. The exposome and periodontal disease: Epidemiologic evaluation of nhanes within smoking classification. *Journal of Clinical Periodontology*, To be submitted 2017.

[121] J Hyman. The importance of assessing confounding and effect modification in research involving periodontal disease and systemic diseases. *Journal of Clinical Periodontology*, Jan 2006.

[122] Vatcheva Kp and Lee M. The effect of ignoring statistical interactions in regression analyses conducted in epidemiologic studies: An example with survival analysis using cox proportional hazards regression model. *Epidemiology: Open Access*, 06(01), 2016.

[123] BL Johnson. *What Are The Human Health Effects Of PCBs?*

[124] Learn about polychlorinated biphenyls (pcbs). *EPA*, Aug 2017.

[125] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016.

[126] T Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 2004.

[127] H D Sesso, J E Buring, W G Christen, T Kurth, C Belanger, J MacFadyen, V Bubes, J E Manson, R J Glynn, J M Gaziano, and et al. Vitamins e and c in the prevention of

cardiovascular disease in men: the physicians' health study ii randomized controlled trial. *Journal of American Medical Association*, 300(18):21232133, Nov 2008.

[128] Office of dietary supplements - vitamin e. *NIH Office of Dietary Supplements*.

[129] N J McKeown. Toluene toxicity. *Background, Pathophysiology, Epidemiology*, Aug 2017.

[130] M Noh et al. Assessment of il-6, il-8 and tnf- levels in the gingival tissue of patients with periodontitis. *Experimental and Therapeutic Medicine*, 6(3):847851, 2013.

[131] J. Walston. Serum antioxidants, inflammation, and total mortality in older women. *American Journal of Epidemiology*, 163(1):1826, Mar 2005.

[132] J A Williams, N Kondo, T Okabe, N Takishita, D M Pilchak, E Koyama, T Ochiai, and D Jensen. Retinoic acid receptors are required for skeletal growth, matrix homeostasis and growth plate function in postnatal mouse. *Developmental Biology*, 328(2):315327, Apr 2009.

[133] H. H. Conaway, E. Persson, M. Halen, S. Granholm, O. Svensson, U. Pettersson, A. Lie, and U. H. Lerner. Retinoids inhibit differentiation of hematopoetic osteoclast progenitors. *The FASEB Journal*, 23(10):35263538, 2009.

[134] M Yamaguchi. Role of carotenoid -cryptoxanthin in bone homeostasis. *Journal of Biomedical Science*, 19(1), 2012.

[135] Agency for toxic substances and disease registry toxicological profile information. *Toxicology and Industrial Health*, 15(8):743746, 1999.

[136] G. W. Bennett, D. L. Ballee, R. C. Hall, J. E. Fahey, W. L. Butts, and J. V. Osmun. Persistence and distribution of chlordane and dieldrin applied as termiticides. *Bulletin of Environmental Contamination and Toxicology*, 11(1):6469, 1974.

[137] W.h. Newsome and J.j. Ryan. Toxaphene and other chlorinated compounds in human milk from northern and southern canada: A comparison. *Chemosphere*, 39(3):519526, 1999.

[138] Toxic substances portal - chlordane. *Centers for Disease Control and Prevention*, Oct 2014.

[139] D.-H. Lee, I.-K. Lee, S.-H. Jin, M. Steffes, and D. R. Jacobs. Association between serum concentrations of persistent organic pollutants and insulin resistance among nondiabetic adults: Results from the national health and nutrition examination survey 1999-2002. *Diabetes Care*, 30(3):622628, 2007.

[140] D.-H. Lee, I.-K. Lee, M. Porta, M. Steffes, and D. R. Jacobs. Relationship between serum concentrations of persistent organic pollutants and the prevalence of metabolic syndrome among non-diabetic adults: results from the national health and nutrition examination survey 19992002. *Diabetologia*, 50(9):18411851, Dec 2007.

[141] M. Porta. A strong dose-response relation between serum concentrations of persistent organic pollutants and diabetes: Results from the national health and nutrition examination survey 1999-2002: Response to lee and others. *Diabetes Care*, 29(11):25672567, 2006.

[142] D A Olsen. Mercury toxicity. *Background, Etiology, Epidemiology*, Aug 2017.

[143] P Grandjean, H Satoh, K Murata, and K Eto. Adverse effects of methylmercury: environmental health research implications. *Environmental Health Perspectives*, 118(8):11371145, Aug 2010.

[144] Kirsten Brandt. Vegetables and fruit in the prevention of chronic age-related diseases. *Molecular Basis of Nutrition and Aging*, page 707722, 2016.

Vita

Appointments

January 2015 - present Statistical Research Coordinator, University of Kentucky

August 2011 - July 2012 Lecturer of Mathematics and Statistics, Eastern Kentucky University

January 2011 - May 2011 Lecturer of Mathematics and Statistics, Murray State University

Degrees

2014 MS in Statistics, University of Kentucky

2010 MS in Mathematics, Murray State University

2009 BS in Mathematics, Murray State University

Workshops & Conferences

2016, 17 Aunual Southern Regional Council on Statistics (SRCOS)

2015, 16, 17 Annual UT-KBRIN Bioinformatic Summit 2015.

2014 19th Summer Institute in Statistical Genetics at The University of Washington. Modules attended: Mixed Models in Quantitative Genetics, MCMC for Genetics, and Advanced Quantitative Genetics.

Awards and Scholarships

2016 Boyd Harshbarger Student Travel Award for travel reimbursement to 2016 Southern Regional Council on Statistics. $500

2014 Workshop Scholarship for module costs and travel reimbursement from Summer Institute in Statistical Genetics at The University of Washington. $2,100

2010 Summer 2010 research grant from Murray State Water Shed Institute. $2000

2010 Research support from Murray State Water Shed Institute. $500

Presentations

2017 Statistical Assistance for (Kentucky) Regional Investigators : 2017 Kentucky Bioinformatics Retreat

2017 Introduction to Shiny : 2017 KBRIN bioinformatics summit

2016 Github and R packages : University of Kentucky Statistics Student Seminar

2015 Shiny Apps and Shiny Servers. Systems Biology : UKy Omics Integration Journal Club.

Applications and R Packages

A REDCap Application which interacts with R and the REDCap database. Alpha.

rFSA R package for running a Feasible Solution Algorithm to find a set of mth order interaction models that are statistically optimal (no one swap can improve the criterion function). Version 1.0.

Statistical Genetics Exploration Application. Beta Version.

An Application for a Mulitcore Feasible Solution Algorithm(FSA) for Finding Interactions . Beta Version.

Manuscripts

Planned Submission Jan 2018 Lambert J, Stromberg AJ, Thompson K. ”A Feasible Solution Algorithm for finding Statistically Significant interactions in Large Datasets.” Journal of Statistical Software.

Planned Submission Jan 2018 Elliott et al., ”Model Feasibility as a Mechanism forUnsupervised Identification ofDomain-Specific Interactions.” International Journal of Statistical Modeling.

Accepted; awaiting publication Black P, et al., "In-program monitoring of student success may be more valuable than predicting success using admissions data." American Journal of Pharmaceutical Education.

Apr 2017 Grulke E, et al., "Size and shape distributions of primary crystallites in titania aggregates." Advanced Powder Technology.

Jan 2014 Slavova, Svetla, Terry L. Bunn, and Joshua W. Lambert. "Drug Overdose Deaths, Hospitalizations, and Emergency Department Visits in Kentucky, 2000-2012.".

2013 Lambert, Joshua W., and Svetla Slavova. "Suicide and Suicide Attempts in Kentucky, 2001-2012.".

Current Support and Grants Under Review

University of Kentucky Healthcare partnership with University of Kentucky Department of Statistics: assist in predictive modeling and data analytics as well as communicate statistical concepts to UKHC senior leadership.

Nathional Multiple Sclerosis Society. Funded March 2017. Identifying Gene or SNP based Interactions in Multiple Sclerosis Data sets.

NIH/NCRR. KY-INBRE: Kentucky Network Biomedical research. Funded May 2014. Statistical Research Coordinator since January 2015.

Research and Analysis Services and Consultation for the Kentucky Justice and Public Safety Cabinet for Needs Assessment and Strategic Planning. January 2017-June 2017