ANALYSIS OF TIME-TO-EVENT DATA, INTERMEDIATE PHENOTYPES, AND SPARSE FACTORS IN THE OPPERA STUDY

Naomi C. Brownstein

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2013

Approved by:

Eric Bair

Jianwen Cai

Michael Kosorok

Gary Slade

Donglin Zeng

ABSTRACT

NAOMI C. BROWNSTEIN: Analysis of Time-to-event Data, Intermediate Phenotypes, and Sparse Factors in the OPPERA Study (Under the direction of Drs. Eric Bair and Jianwen Cai)

Motivated by the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) project, a large study of temporomandibular disorders (TMD), this dissertation develops statistical methods applicable to three facets of chronic pain.

First, we propose a method for parameter estimation in survival models with missing censoring indicators. These result because conducting multiple invasive examinations for incidence on all participants in large prospective studies is infeasible. We estimate the probability of being an incident case for those lacking a gold standard" examination using logistic regression. Multiple imputations of case status for each missing examination are generated using these estimated probabilities. Imputed and observed data are combined in Cox models to estimate the incidence rate and associations with putative risk factors. The variance is estimated using multiple imputation. Our method performs as well as or better than competing methods and highlighted new discoveries for OPPERA.

Secondly, we propose a general method to analyze secondary phenotypes and apply it to the OPPERA baseline case-control study. Traditional case-control genetic association studies examine relationships between case-control status and one or more covariates. Investigators now commonly study additional phenotypes and their association with the original covariates as secondary aims. Assessing these associations is statistically challenging, as participants do not form a random sample from the population of

interest. Standard methods may be biased and lack coverage and power. Utilizing inverse probability weighting and bootstrapping for standard error estimation, our method performs as well as competitors when they are applicable and provides promising results for outcomes to which other methods do not apply.

Third, we propose a method for sparse factor analysis. Psychometric studies frequently measure numerous variables that may be noisy manifestations of a few underlying constructs. Aims include identifying these latent variables and their relationship to the observed variables and reducing the data to a few key variables that explain the majority of variance. While variable reduction methods exist for principal component analysis, none have been proposed to date for factor analysis. Our method retains predictive accuracy for many thresholds in simulations while providing sparse loadings. Competing methods had less predictive accuracy or less sparsity.

To my little sister, Shira Brownstein, in loving memory. Both her presence and absence have shaped my progress throughout the program and continue to inspire me to work hard and never give up.

ACKNOWLEDGMENTS

People commonly say that it takes a village to raise a child. It also takes a villiage to complete a PhD. There are numerous individuals to whom I owe gratitude in regards to this endeavor. I apologize in advance to any individuals I have inadvertently omitted.

First and foremost, I am deeply appreciative of my co-advisors, Drs. Eric Bair and Jianwen Cai. Both have shaped me immeasurably throughout my time at UNC, providing not only technical insight on the research, but serving as caring mentors and encouraging me to work hard and develop to my full potential. This dissertation would not have been completed, especially not in this time frame, without their dedication and support. Secondly, I am grateful for the committee members, namely Drs. Michael Kosorok, Gary Slade and Donglin Zeng, for taking the time to meet with me and provide helpful questions and comments that strengthened the research. I also appreciate the staff at UNC, such as Melissa Hobgood, Veronica Stallings, Christine Kantner, and especially Mae Beale and Dean Barbara Rimer, without whom I likely would not have finished this degree.

Peers have been invaluable, as we've navigated through this journey together. Thank you to Steve Hoberman, Lan Liu, Amy Richardson, Joe Rigdon, and Eugene Urrutia, among others, for studying together in courses and qualifying exams, creating a cooperative atmosphere in which to grow as statisticians, and for your friendship. Thank you to those who came or finished beforehand and offered helpful advice, such as Eric Daza, Beth Horton, Diana Lam, Che Smith, and Alison Wise, and Drs. John Bainbridge, Jennifer Clark, Annie Green Howard, Suprateek Kundu, Dustin Long,

Leann Long, and Greg Mayhew.

My graduate study was made financially possible by generous support from the NSF Graduate Research Fellowship Program grant #0646083 and NIH/NIEHS T32ES007018, led by Dr. Amy Herring. I would like to acknowledge and thank the principal investigators of the OPPERA study for providing the data which was analyzed in this dissertation. Specifically, this includes Drs. William Maixner, Luda Diatchenko, Bruce Weir, Richard Ohrbach, Roger Fillingim, Joel Greenspan, and Ronald Dubner. The OPPERA study was supported by NIH/NIDCR grant U01DE017018.

Finally, I am grateful for the vast emotional support of family and friends. Thank you to my parents, sisters, grandparents, and extended family, for being here for me throughout the many ups and downs of this time. I especially appreciate the company of and opportunity to become closer with the Reed family. Thank you to the many friends who offered advice, comfort and joy when it was most needed including, but not limited to Jessica Meed and Bronco Suzuki, Charlotte and Michael Hall, Laenja and Matt Brackett, Elizabeth Clendinning, Paul Baek, Anthony Kuhns, Correai Moore, Jonathan Rowell, Alyssa Manning, and many others on the UNC Ballroom Dance Team and in UNC Hillel. Last but not least, thank you to Jesse Haskins for supporting me, especially around the qualifying exam and during our year of long distance while I completed the dissertation. I look forward to beginning the next chapter of our lives together.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Time-to-event analyses are frequently conducted in medicine, actuarial science, and numerous other fields of applied science. Actuaries study the time until death of individuals for the purpose of calculating and offering fair life insurance rates. In clinical trials, researchers note whether and when participants experience the event of interest and compare the times between the treatment and control groups. Similarly, in cohort studies, researchers compare survival times and want to know if survival time is related to a risk factor of interest. Additionally, it may be of interest to study the hazard of death over time.

There is a well-developed set of standard time-to-event analysis methods implementable in standard statistical software packages. Logrank tests allow testing of differences of survival times between a finite number of groups. The survival distribution may be eassily estimated non-parametrically and plotted using SAS or R. Semi-parametric methods, such as the Cox proportional hazards model, allow robust estimation of the hazard function in the presence of covariates. Yet, these methods require knowledge of both the event time and status for all individuals. The event status may not be known for all individuals, especially when one is interested in studying death due to a particular cause.

Additionally, current methods do not work for secondary time-to-event outcomes in case-control studies, in which study participants are sampled based on the primary outcome, because the study does not constitute a simple random sample of the population

of interest. In fact, arbitrary analysis of one or more secondary phenotypes is a non-trivial problem and has spawned a great deal of research in recent years.

Often, the outcome of interest may be difficult to ascertain. For example, an autopsy by expert medical examiner may be needed to determine if a death was due to myocardial infarction, a cancer tumor, or other factors. In order to mitigate this problem, some studies employ delayed event adjudication. That is, possible cases are identified using simple, but possibly error-prone methods. Then, one or more experts examines the possible cases using a more accurate, but also more costly and time-consuming method to determine the true event status. For example, a specialized dental examination is required for accurately diagnosing temporomandibular disorders (TMD). It is impractical to subject large number of subjects to such an examination, especially if they are unlikely to have the condition. Instead, the "gold standard" examination is performed only on subjects who screen positive on a simpler assessment, such as a questionnaire designed to measure recent orofacial pain. However, some subjects do not receive the "gold standard" examination due to inability or unwillingness to attend research centers. A time-to-event analysis would then have some subjects with self-reported symptoms but missing censoring indicators. This setting presents statistical challenges, which require care in order to avoid bias and maintain efficiency.

In other settings, cause of death may be unknown or death certificates may be missing entirely. This frequently arises when there are multiple failure types. In oncology studies, for example, researchers may want to differentiate between deaths due to cancer and deaths due to car accidents or other unrelated causes. For such studies, a death certificate is insufficient to classify a subject. Moreover, it may be impossible to determine if a death occurred if a subject dies abroad or national death registries are incomplete.

While there have been a few methods developed in this area, each has drawbacks.

There is a clear need for new methodology to handle this challenging situation. In the first paper, we propose a method for parameter estimation in the case of missing censoring indicators. In the second paper, we seek an unbiased and efficient method to model the relationship between a genotype and a secondary phenotype in case-control studies. The third paper extends the framework of the first paper to allow for Cox regression with both missing covariates and missing censoring indicators.

CHAPTER 2: LITERATURE REVIEW

Our methods, detailed in chapters 3-5, rely on basic knowledge of survival analysis, missing data, imputation, bootstrapping, joint-modeling, case-control genetic studies and factor analysis. This chapter reviews the literature pertaining to these fields.

2.1 Survival Analysis

Actuaries, medical experts, and statisticians alike are frequently confronted with failure time data. One important quantity of interest is the survival function. The nonparametric maximum likelihood estimate of the survival distribution is given by (60). Nonparametric estimation in the accelerated failure time model is discussed in (72). The shape of the survival function is instructive in studying the progression of the event of interest over time.

Yet, it is often of interest to model the relationship between the failure times and a set of covariates. A parametric model with one covariate assuming that the failure time has an exponential distribution is proposed in (36). Unfortunately, the stringent distributional assumption and limit of one covariate prevent this method from being useful in most situations. Instead, the classic semi-parametric model of the hazard function in the presence of covariates is defined in (10). The field of survival analysis in the past forty years has been centered around this model. In fact, in prospective cohort studies of risk factors for disease, hazards ratios are beneficial because they

approximate the incident rate ratio. The pervasiveness of this model is evident in, for example, (40), a review of developments in survival analysis in clinical trials through the turn of the century.

2.1.1 Cox Proportional Hazards Model

One of the most popular statistical tools today, the Cox model combines the ability to handle covariates with the flexibility resulting from leaving the baseline hazard unspecified. Cox defines a partial likelihood, which allows for maximum likelihood estimation of the regression parameters, and extends results for bivariate life table data. Further details on the partial likelihood are provided in (11).

The Fisher information matrix for the regression parameter from the standard likelihood is calculated in (26) and shown under broad conditions to be asymptotically equal to the information based on the partial likelihood. Thus, the Cox likelihood is asymptotically efficient. Efron also estimates the hazard rate and connects it to the Kaplan-Meier estimate of the survival function. Accelerated failure time model and inference and parameter estimation in the Cox model is discussed in (59) as well as extensions to multivariate failure data, time-dependent covariates, and case control studies. A heuristic method for computing the asymptotic variance of the estimated survival function is proposed in (68) and used it to construct confidence intervals, which are shown in simulations to have adequate coverage.

The applicability of Cox models continued to grow. Methodology were generalized to competing risks and multivariate outcomes. The framework for competing risks is detailed in (81), (50) and (94).

Another important question is how robust the Cox model is to mis-measured or missing data and what modifications are appropriate in the presence of such data. Effects of covariate measurement error are investigated in (93) and (52). Cox models

with missing data are discussed in section 2.4. Specifically, the setting of uncertain outcomes are discussed in section 2.4.2.

2.2 Bootstrapping

Bootstrapping is an estimation method that may be applied in a wide variety of situations with or without missing data. The goal is to estimate a statistical quantity and its variability. Bootstrapping is especially useful for quantities whose standard error is difficult or impossible to calculate mathematically. Over the past 30 years, the original method has evolved dramatically, with new methods applicable to frequentist, Bayesian, parametric, nonparametric, and semiparametric situations alike.

The nonparametric bootstrap was originally proposed in (27) as a modification of the jackknife. The first step is to find the empirical distribution function of the data. Next, one creates a large number of boostrap samples by sampling with replacement from this empirical distribution function. Third, one combines the results from each of the replications. The estimate of the quantity of interest is the mean of the R estimates from the replicates and the estimated standard error is the standard deviation of the estimates from the replicates. Details may be found in (27) and (29).

This simple procedure, available in standard statistical software, originally allowed estimation of the parameter of interest and of bias, but now also routinely provides several confidence intervals. Asymptotic properties, including second order accuracy, are discussed in (111), (5) and (112). Literature on standard errors and confidence interval estimation is discussed in more detail in section 2.2.1.

In its classical form, bootstrapping is a nonparametric frequentist method, though it has been adapted to the parametric and Bayesian frameworks as well. One example of parametric bootstrapping is found in (29). Instead of substituting the empirical

distribution for the unknown sampling distribution, Efron substitutes a normal distribution with mean and variance equal to the sample mean and sample variance. This procedure is called the normal smoothed bootstrap. (102) first described the Bayesian bootstrap. Weighted bootstrapping, which is applicable to hypothesis testing, is discussed in (62). Bootstrapping has emerged as a useful tool for practitioners and the foundation of numerous papers on its theoretical properties, scope, and limitations.

In his silver anniversary review paper, (31) stresses that boostrapping is founded on the "plug-in principle", which he defines as the act of simply "plugging in" the empirical distribution function for the unknown distribution function. He notes that while the literature as of 2003, (19), indicates that this is generally valid to the second order, there are notable situations where it can be problematic. According to a similar review paper by (14), it is the simplicity of the bootstrap and its connection with other methods, such as the jackknife, that make it both widely utilized in applications and extensively studied from a theoretical standpoint.

2.2.1 Standard Error and Confidence Intervals

One important benefit of bootstrap methods is the ability to calculate standard errors and confidence intervals. For many estimators of interest in practice, standard errors are, at best, difficult to calculate, or worse, impossible to evaluate analytically. Moreover, even when the standard error can be found, standard confidence intervals based on the central limit theorem may be inadequate for statistics whose distribution are non-normal. On the other hand, according to (32), the bootstrap can always provide numerical standard error estimates. A simple example of the bootstrapping in the Cox proportional hazards model can be found in (32).

Three methods of bootstrap confidence intervals are discussed in the literature,

namely percentile, bias corrected and acclerated (BCa), approximate bootstrap confidence intervals (ABC) and bootstrap 't'. The percentile method simply reports the estimated percentiles of the bootstrap distribution. A quick theoretical justification for percentile intervals is provided in (28). BCa intervals are motivated by the assumption that a simple monotone transformation of the parameter of interest is normally distributed with slight bias with respect to the transformed mean and a variance that may not be constant for all possible parameter values. The user does not need to specify the transformation. In order for BCa confidence intervals to be accurate, it is only required that such a transformation exists. Monte Carlo sampling yields adequate estimates of the standard error in about 50-200 replications for most statistics and adequate BCa confidence intervals in about 1000 replications (32, 19). The bootstrap 't' interval is conceptually simpler than BCa intervals. However, unlike percentile intervals, they are not transformation invariant and can result in confidence intervals that are much too wide. ABC intervals do not require Monte Carlo simulations at all, relying instead on an analytic approximation to the BCa interval.

2.3 Missing Data: General Theory and Methods

The problem of missing data is nearly ubiquitous in practical statistical problems. Consequently, there is a rich body of research on analysis with missing data, including methods such as the EM algorithm, single imputation, and multiple imputation. There are also a wide variety of ad-hoc techniques that may introduce severe bias. This section summarizes the literature on missing data and proper data analyses.

In the classic paper, (101) defines three missing data mechanisms and ignorability. Before proceeding, necessary definitions from Rubin's work are introduced. Additional details may be found in (70). The missing data mechanism refers to the probability

of having missing data. Data are called missing completely at random (MCAR) if the probability of having a missing value is independent of the data, both observed and missing. Data are called missing at random (MAR) if the probability of having a missing value is independent of the missing values, but may depend on the observed values or on observed covariates. Under MAR or MCAR, the missing data mechanism is said to be ignorable. If the probability of having a missing value depends on that unobserved value, then the data is missing not at random (MNAR) and the mechanism is non-ignorable.

When confronted with missing data, data users face a dilemma. They may know that, in theory, disregarding observations that are not fully observed is inefficient, at best. The loss in efficiency increases with the proportion of missing data. It may bias inference and lead to misleading conclusions. Yet, most software is not equipped to handle missing observations. Typically, observations with any missing values are dropped. This can be problematic in studies with a large number of covariates of interest. Subjects with one or more of the covariates missing would be excluded. Denoted a "complete-case" analysis, the results may be severely biased when the missing data mechanism is non-ignorable. Similarly, "available-case" analyses, which use all cases with observed values for at least one of the variables under consideration, may be invalid. Mean-imputation, in which missing values are imputed by the mean of observed values, may also introduce bias, (103). While it may be tempting to use these or other more convenient approaches, it is usually not appropriate. There is a rich set of methodology for handling missing data without introducing bias and maintaining as much efficiency as possible. A selection of relevant methods are included below.

2.3.1 The EM Algorithm

One common method for handling missing data is the Expectation-Maximization (EM) Algorithm. There are three steps: first, finding the expected value of the full log-likelihood given the observed data and current parameter estimate, second maximizing the expectation to update the parameter estimates, and third, iterating until convergence. The method was formally introduced in (16) with details of the E-step and M-step. The authors demonstrate why the EM algorithm works, that is, the fact that the likelihood always increases with each iteration. They also give examples and suggest the ability to generalize the EM algorithm in a Bayesian framework.

The EM-algorithm in some form is instrumental in most papers that develop new methods for analysis of data with missingness. One noteworthy example is in the context of logistic regression with uncertain outcomes. The EM-algorithm and known or estimated values of sensitivity and specificity are used in (74) to improve upon parameter estimates in standard logistic regression. Section 2.4 discusses additional examples of methods that employ the EM-algorithm for missing data in survival analysis.

2.3.2 Multiple Imputation

Multiple imputation, discussed in section 2.3, is a Bayesian method for analyzing data with some degree of missingness. It was first presented as a way to handle non-response in sample surveys. The framework is summarized in (103) as containing a database constructor often separate from the person who will analyze data. He states that multiple imputation is a statistically valid method within this framework. Multiple imputation definitions and equations are reviewed. He defines proper multiple imputation and urges users to include available covariate information in the imputation.

Finally, Rubin compares multiple imputation to competing methods, i.e. single imputation, jackknife, bootstrapping.

Missing data mechanisms are discussed in (155) from frequentist and Bayesian points of view as well as the consequent challenges under various assumptions and appropriate missing data analysis methods. Imputation in general is explored in (155), specifically two multiple imputation methods – propensity score and predictive model – under monotone missingness, and one method of multiple imputation under non-monotone missingness utilizing Markov Chain Monte Carlo (MCMC). Readily vailable multiple imputation software procedures, such as PROC MI and PROC MIANALYZE in SAS, frequently assume data are MAR.

Often, especially for binary data, values generated with multiple imputation are rounded. Yet, rounding can cause excessive bias, as shown in (51). One solution is to keep the raw imputed values, as they introduce less bias than the rounded values. They urge instead to posit a more appropriate posterior, for example to refer to Rubin's detailed instructions for how to impute missing binary data.

2.3.3 Bootstrapping and Missing Data

For missing data problems, (30) provides a comparison of three bootstrapping methods and multiple imputation. Non-parametric bootstrapping, the simplest form of bootstrapping, requires no knowledge of the missing data mechanism. It may be applied to virtually any reasonable estimator. However, nonparametric bootstrapping may be too slow or infeasible to use in practice due to the comparatively large number of replications required. This is especially true when BCa confidence intervals are of utmost interest.

Full mechanism bootstrapping is less computationally intensive, but requires modeling

of the missing data mechanism. Thus, it is most appropriate for non-ignorable missing data, when this is required anyway. The multiple-imputation bootstrap begins with the Bayesian bootstrap and calculates confidence intervals via the ABC method. It is most appropriate for Bayesian settings, because it requires knowledge of the conditional density of the full data given the observed data. Note that the requirement to know the conditional density is less stringent than the requirement of full-mechanism bootstrapping to know the missing data mechanism. For censored surival data, full mechanism and nonparametric bootstrapping are identical (29, 30).

An important type of partially incomplete data is censored failure-time data. Under the assumption of random censoring, (29) argues that bootstrapping is valid for survival data. Similarly, it is straightforward to use bootstrapping to estimate the survival function and its standard error.

2.4 Missing Data Methods and Survival Analysis

This section discusses specific developments in missing data methods for time-to-event data. As discussed previously, survival analysis already has incomplete data in the sense that the failure time is unknown for censored individuals. Most methods assume that data are censored randomly. In the literature, this is termed as ignorable censoring. When this assumption is not tenable, censoring is termed non-ignorable and adjustments must be made.

When modeling time-to-event data with covariates, there are additional types of missing data that may arise. Most commonly, some subjects may have one or more unobserved covariates. A less studied situation is when the censoring indicator is missing for a subset of the observations. While it is possible for a single dataset to have missing covariates and missing censoring indicators, no papers on that problem

have been found to date.

The following subsections detail, respectively, the literature for survival analysis with missing covariates and missing censoring indicator models.

2.4.1 Cox Models with Missing Covariates

There have been numerous papers on modeling proportional hazards with missing covariates. Estimating equations are propposed in (65) to yield what they call an "approximate partial likelihood estimator" (APLE), which is consistent and asymptotically normal under MCAR covariates. The assumption is relaxed to MAR by (89) and impute the covariates using the conditional expectation based on observed information. Through simulations, they demonstrate the superiority of their method to that of Lin and Ying in terms of efficiency under MCAR and consistency under MAR. A nonparametric model for estimation of relative risk with the missing covariates is given in (156) based on additional auxiliary covariates. An EM-type algorithm for missing covariates in Cox models is detailed in (49). Additional papers include the inverse selection probability weighted estimator of (139), among others.

2.4.2 Cox Models with Missing Censoring Indicators

Missing censoring indicators frequently arise when there are multiple failure types. Investigators may easily record the mortality of all subjects, but it may be extremely difficult or costly to find out exactly why each subject died. Consequently, there has been much more work done on missing censoring indicators in the context of competing risks than when there is only one type of failure. Over the past three decades, authors have produced a variety of methods of estimating the hazard and survival function

for the missing censoring indicator model. The logrank test has been generalized for missing censoring indicators. Yet, there have been fewer papers on the more difficult problem of Cox regression.

First, (20) uses the EM algorithm to estimate the survival function under various types of partially observed competing risk data. Observations may be either fully observed, censored, or observed failures with unobserved causes. He estimates the information matrix and provide clear examples of the method applied in practice. The survival function is estimated non-parametrically in (96) when death status is known but its cause is uncertain for some subjects, while the Kaplan-Meier estimate may be biased for such data. Causes of failure are said to be masked when it is unknown but at least one type of cause may be eliminated from consideration. For competing risks data with masked causes of failure, (39) estimate the survival function.

A modified logrank test for competing risk data for which cause of failure is unknown and missing at random, (44). A slight modification to their partial likelihood is proposed in (17) to increase the information and improve the test. Multiple imputation of the cause of failure and a corresponding asymptotically valid modified log-rank test is detailed in (133).

For standard survival data, there have been a number of methods proposed for estimating the survival function under the missing censoring indicator model. An asymptotically efficient estimator of the survival function under missing at random, or more generally, "coarsening at random" censoring indicators, is proposed by (136). Namely, they use the nonparametric MLE of the survival function based on reduced data produced by a discretization of the failure time. A comparable alternative is proposed by (123) that requires less computational resources and time. Kernel estimation is employed in (124) to provide a weakly convergent estimator of the survival function. As in the case with fully observed censoring indiators, kernel estimation depends on

the choice of bandwidth sequence. It follows that clever estimation of the bandwidth would improve the kernel estimation performance. Three methods of bootstrapping the bandwidth are compared to each other and to existing methods, (129). Semiparametric random censorship models for hazard function estimatiors, both nonparametric and semiparametric, are given in (128). In addition, they compute and minimize the asymptotic mean squared error to find the optimal bandwidth. An augmented inverse probability weighted estimator of the survival function under MAR is proposed in (127) and shown that it is robust even under partial model misspecification. Improving the survival function estimate in the missing censoring indicator model via multiple imputation is discussed in (125) and (126).

An important research question is how Cox models are affected when case status is uncertain. Competing risks proportional hazards regression models are proposed in (45) using estimating equations assuming that observations are known either to be censored or to have failed by some cause. They assume proportional hazards for each failure type and between the two failure types. For the linear transformation model, An augmented inverse probability weighted estimator and algorithm are proposed in (41) along with asymptotic and double robustness properties.

Time-to-event data which may not be classified and confirmed is modeled in (8) by weighting each subject by their estimated probability of being a true case. Their methods include estimation of the survival function, Cox models, and a logrank test. They propose an asymptotic variance expression which may be used as an approximation for finite samples. Alternatively, they suggest estimating the variance by bootstrapping. Empirical processes are employed to prove consistency and weak convergence of the estimators to Gaussian processes.

For the usual survival setting with only one cause of failure, (43) estimate the survival function by modifying and combining estimates of the hazard function, discuss

estimation in the Cox model and prove consistency and asymptotic normality of the estimators. The survival function is also estimated without covariates and Cox proportional hazards models are proposed in (78) by modifying the Nelson-Aelen estimator and taking the product-integral. The advantage of their method is that it does not require discretization of the failure times or estimation of the unconditional probability of having an observed censoring indicator. Simulations demonstrate that the (78) is superior to previous estimators, such as the earlier version of (43). Yet, the methods of (78) and (43) both depend on the MCAR assumption, which they state that it is non-trivial to relax. Parameter estimation in the Cox model under the assumptions of proportional censorship and MCAR censoring indicators is conducted in (122).

Weights are recommended for each potential event, of which individuals may have more than one, along with a weighted partial likelihood, (116). That is, he suggests screening out observations that are unlikely to be true events, and weighting each observation by the probability that the subject had first failed at the observed time. The weights may be known or estimated using auxiliary covariates. Taking this uncertainty into account reduces bias and increases power. However, estimating the weights depends upon the presence of a normally distributed diagnostic variable and either knowing or having experts guess the relative frequency of true endpoints to false endpoints. In addition, the variance of the method is mentioned but not described in detail.

Adjusting a proportional hazards model based on discrete survival times and measurement error of case status is discussed in (79). They show that mis-measured outcomes can bias the usual proportional hazards model parameter estimates toward the null and poor coverage probability for confidence intervals, but this can be mitigated by incorporating information about sensitivity and specificity.

Estimators for the regression parameters in the Cox proportional hazards model with missing censoring indicators are derived in (6) using the EM algorithm and

their consistency is established under basic regularity conditions. The authors then estimate the non-missing rate of censoring indicators with the kernel estimator and use it to estimate the information matrix. Their approach depends on the assumption of piecewise constant hazard functions and proportional hazards for not only the event time, but also for the censoring time.

Multiple imputation is also discussed for repeated failure times where the event is known to have occurred at a certain time but the event type is missing (107). The method is compared to multiple imputation under the Cox proportional hazards model without repeated events.

2.5 Genetics and Case-Control Data

Investigators are frequently interested in reusing case-control data to evaluate associations between measured risk factors and outcomes other than the outcome to define case status. These other outcomes are often referred to as secondary phenotypes, which may be associated with the primary disease used to define case status. Utilizing one study to investigate more than one outcome is important for financial reasons, as the process of attaining genetic information and other putative risk factors for all participants is time-consuming and expensive. While logistic regression is known to be invariant to the sampling method for primary analyses, this does not hold for secondary phenotype analyses. Simple logistic and linear regression may be severely biased. Restricting analyses to cases only or controls only results in decreased efficiency. These methods are considered "naive" methods. There has been a great deal of recent interest in developing unbiased and efficient methodology for analyzing secondary phenotypes in case-control studies.

Two methods for analyzing associations are proposed in (97). The first is to use a

standard logistic regression model with a covariate adjusting for original case status. The second is to use stratum-weighted logistic regression, i.e. logistic regression in which cases and controls are weighted by the reciprocal of their sampling fractions. For population studies, it is often easier to estimate the ratio of the sampling fractions than to estimate the sampling fractions for cases and controls separately. Hence, (97) proposes using a unit weight for the cases and the ratio of the sampling fractions for controls. In a nested case-control study, the ratio is the number of cases in the whole study divided by number of controls.

Simulations are conducted in (82) to estimate size and power for the naive methods and the inverse probability weighting (IPW) method and dummy variable adjustment method in (97). They conclude that IPW has adequate type I error but has increased variance and sometimes decreased power compared to the naive approaches. Note, however, that the simple unweighted analysis and the analyses that adjust for case status with a binary variable or restrict to either cases or controls have slightly inflated type I error when the primary disease is not rare and both the genetic profile and secondary phenotype are related to the primary disease. There are also situations in which the IPW method is more efficient than the method that adjusts for case-control status with an indicator variable for case status.

A rigorous likelihood based approach is derived in (66). However, (63) points out that (66) assumes no interaction between the covariate information (genotypes) and the secondary phenotype and that in the case of interaction in modeling the probability of having the primary disease, the results may be biased. They propose an adaptive weighting parameter estimate motivated by a Bayesian shrinkage estimator for gene-environment interaction when the disease is rare.

The bias may be corrected by solving iterated non-linear equations relating regression coefficients to estimates of prevalence of the disease and phenotype from the literature

and using bootstrapping to calculate confidence intervals via the percentile method, (140). Their method assumes that the secondary phenotype may be modeled by logistic regression given a set of binary covariates and the primary disease may be model by logistic regression given the covariates and the secondary phenotype but no interaction. They describe the IPW method in (97) and extend it to retrospective case-control studies using an estimate of disease prevalence. They also adapt their bias correction method to the frequency-matched case-control study design. Their method produced narrower confidence intervals than the IPW method.

In their second paper, (141) conduct simulations, which show that the method in (140) has adequate size and is more powerful than the naive logistic regression approaches. A joint model based on a Gaussian copula is proposed by (47) for one or more secondary phenotypes in the exponential family and the disease of interest given the covariate, i.e. genotype. Their simulations confirm that their method is more powerful than the IPW method. Unlike other methods, this one allows joint modeling of multiple secondary phenotypes that is more powerful than multiple univariate analyses.

2.6 Sparse Factor Analysis

Latent variable models are frequently used in statistics. In certain situations, it is reasonable to assume that observed variables are random variations of one or more unobserved random variables. Within the field of psychometrics, studies often consist of a large number of correlated measured variables, which may be manifestations of a smaller number of underlying constructs. Factor analysis is commonly used for this purpose. In factor analysis models, the observed data is assumed to be a linear combination of the unobserved factors, plus random error. See (9) for an excellent description and illustration of some simple factor analysis models. A more

theoretical description of the problem of factor analysis is provided in (3) along with a demonstration that it may fall into the Bayesian framework.

Factor analysis may be either exploratory or confirmatory in nature. Exploratory factor analysis (EFA) aims to discover underlying relationships between the observed variables and is more commonly used (91). Confirmatory factor analysis (CFA) begins with an a priori structure and tests hypotheses about those structures using the factor analysis setting. Researchers such as (137, 55) have debated quite heatedly about when each type of factor analysis is appropriate.

Principal component analysis (PCA) is more frequently used in statistics than is factor analysis. In principal component analysis, instead of assuming that there are unobserved variables driving the observed data, it is desired to extract linear combinations of the observed variables. The first linear combination, or component, explains the largest amount of variability in the data. The second component explains the next largest amount of variance, and so on. Traditionally, principal component analysis extracts the same number of components as there are observed variables. However, since data reduction is commonly a goal of PCA, usually the analyst selects a smaller number of components that explains a sufficient amount of variance.

It is desired to have sparse loadings, meaning that many of the loadings are equal to zero. Unfortunately, neither principal component analysis nor factor analysis produce sparse loadings. This can be problematic, especially in genetic settings, where the number of variables present in the study can be very large. Two methods (157, 147) have recently been proposed for sparse principal component analysis. (147) uses a penalized matrix decomposition to produce more sparse results than traditional principal component analysis.

The aforementioned sparse principal component analysis methods do not immediately generalize to factor analysis models. To date, the author of this dissertation has not

found any articles on sparse factor analysis.

CHAPTER 3: PARAMETER ESTIMATION IN COX PROPORTIONAL HAZARD MODELS WITH MISSING CENSORING INDICATORS

3.1 Introduction

Time-to-event analyses are frequently conducted in medicine, actuarial science, and numerous other fields of applied science. There is a well-developed set of survival analysis methods implemented in standard software. Semi-parametric methods, such as the Cox proportional hazards model, allow robust estimation of the effects of covariates on the hazard function. However, these methods require the analyst to know the censoring status of each participant, which may not always be available.

In some cases the outcome of interest may be difficult to ascertain. For example, in oncology studies, researchers may want to differentiate between deaths due to cancer and deaths due to car accidents or other unrelated causes. Investigators may easily record the mortality of all subjects, but it may be extremely difficult or costly to find out exactly why each subject died. One possible solution to this problem is delayed event adjudication (8). This means that possible cases are not identified immediately but screened using simple methods that may have poor sensitivity or specificity. Later, the screened candidate cases are re-examined using a more precise, but also more costly and time-consuming, method to determine the true event status.

The study that motivates our work is Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA), a prospective cohort study to identify risk factors for the onset of temporomandibular disorders (TMD). Each (initially TMD-free) OPPERA

study participant was followed for a median of 2.8 years to identify cases of first-onset TMD. However, it was impractical to perform a physical examination on every participant. It would also have been inefficient given that most study participants did not develop the condition. Instead, this "gold standard" examination was performed only on participants who screened positively on a quarterly screening questionnaire that was designed to assess recent orofacial pain (1). However, some participants who screened positively were lost to follow-up before receiving the "gold standard" examination. Thus a time-to-event analysis would have some participants with missing censoring indicators.

Cook and Kosorok (8) estimate parameters in Cox proportional hazard models with missing censoring indicators by weighting observations according to their probability of being a true case. They show that the estimators are consistent and asymptotically normally distributed. However, the standard error of their proposed estimate cannot be easily obtained using existing software without bootstrapping. For the OPPERA data, a separate Cox model was calculated for each putative risk factor of interest, including approximately three thousand genetic markers. Consequently, applying this method to the OPPERA genetic data would be computationally intractable.

In the OPPERA study, the likelihood that a participant who screened positively was examined was associated with demographic variables such as gender, race, or socioeconomic status (1). This indicated that the censoring indicators in the OPPERA study were not missing completely at random (MCAR). Application of models that assume MCAR censoring indicators may result in biased estimates of hazard ratios for covariates of interest. More importantly, a participant's responses to their screening questions are predictive of whether or not they are an incident case of TMD. This setting presents statistical challenges, which require care in order to avoid bias and maintain efficiency. Additionally, incidence rate estimates are desired, and none of the

methods currently available allow for estimation of the incidence rate. There is a clear need for new methodology to effectively answer the research questions of the OPPERA study.

In this paper, we propose a method for parameter and variance estimation in Cox regression models with missing censoring indicators. The motivating data set is introduced in section 3.2. We describe our method in section 3.3. In section 3.4, we report the results of simulations. Finally, in section 3.5 we apply our method to the OPPERA study. We conclude with a discussion in section 3.6.

3.2 Motivating Data Set: The OPPERA Study

OPPERA is a prospective cohort study designed to identify risk factors for first-onset TMD. A total of 3,263 initially TMD-free subjects were recruited at four study sites between 2006 and 2008. TMD status was confirmed by physical examination of the jaw joints and muscles using the Research Diagnostic Criteria for TMD (25), which is the gold standard for diagnosing TMD.

Upon enrollment in the study, each OPPERA participant was evaluated for a wide variety of possible risk factors for TMD, including psychological distress, previous history of painful conditions, and sensitivity to experimental pain. For a brief overview of the risk factors of interest in the OPPERA study, see Section 6 in the Web Appendix. See Ohrbach et al. (86), Fillingim et al. (37), Greenspan et al. (46), and Maixner et al. (76) for a complete description of the baseline measures that were collected in OPPERA.

After enrollment, each participant was asked to complete questionnaires to evaluate recent orofacial pain once every three months. These questionnaires (hereafter referred to as "screeners") evaluated the frequency and severity of pain in the orofacial region during the previous three months. The purpose of the screener was to identify participants who were likely to have recently developed TMD. For a complete description of the

screener, see Slade et al. (114). Participants who screened positively were asked to undergo a follow-up physical examination by a clinical expert to diagnose presence or absence of TMD.

Of the 3,263 subjects, 2,737 filled out at least 1 screener, and the remaining 521 did not fill out any screeners. The total number of screeners was 26,666. There were 717 positive screeners, 486 (about 68%) of which were followed by a clinical examination. As reported in Bair et al. (1), case classifications made by one examiner (hereafter, "Examiner #4") were deemed unreliable because the examiner diagnosed a much higher percentage of individuals with TMD compared to other examiners. We therefore set all of Examiner #4's physical examination findings to be missing and imputed them using the methods in this paper. This left 404 positive screeners (56%) resulting in valid clinical exams. On the individual level, after setting the exams from Examiner #4 to missing, there were 404 people who had one positive screener, and 114 people who had two or more positive quarterly screening questionnaires.

3.3 Model

3.3.1 Notation and Assumptions

Assume there are n independent participants. For each participant i ($i = 1, \ldots, n$), let C_i and T_i denote the potential times until censoring and failure, respectively, let $V_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, $N_i(t) = I(T_i < t)$. Let Z_i a $p \times 1$ vector of covariates measured at baseline and let X_i be a $q \times 1$ vector of covariates measured at the time of the putative event. We assume the hazard for participant i follows a Cox proportional hazards model

$$\lambda(t|z_i) = \lambda_0(t) \exp(\beta' z_i) \tag{3.1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function. Let ξ_i denote the indicator that Δ_i is observed and $\sigma_i = \xi_i \Delta_i$. We observe (V_i, ξ_i, σ_i) for $i = 1, \ldots, n$.

In the OPPERA study, V_i is the length of time for participant i between enrollment in the study and either of two events

(a) a screener which resulted in a diagnosis of incident TMD

(b) the last-completed screener before loss-to-follow-up.

If participant i either screened positively and subsequently was diagnosed with TMD, then $\Delta_i = 1$. If participant i either screened negatively on the last quarterly screener before loss-to-follow up or screened positively and was diagnosed to be free of TMD, then $\Delta_i = 0$. If participant i screened positively on the last screener but was not examined, then Δ_i is missing and $\xi_i = 0$. The putative risk factors for TMD that were assessed at enrollment are denoted by the vector Z_i. Responses to the screener for participant i at time V_i are denoted by the vector X_i. For OPPERA, we also define $Q_i = 1$ if participant i screens positively on the last screener before either a positive diagnosis of TMD or loss to follow-up and $Q_i = 0$ otherwise.

We assume the censoring indicators are missing at random (MAR) as follows:

$$P(\xi_i = 0 | X_i, V_i, \Delta_i, Q_i = 1) = P(\xi_i = 0 | X_i, V_i, Q_i = 1) \tag{3.2}$$

In other words, the probability of having a missing censoring indicator may depend on measured factors, but it does not depend on whether or not an event occurred. We will denote the probability in (3.2) by $\rho(X_i, V_i)$.

3.3.2 Estimating Event Probabilities

We model the probability that participant i with a missing censoring indicator is a case by a logistic regression model based on X_i and V_i:

$$\begin{aligned} & P(\Delta_i = 1 | X_i, V_i, Z_i, \xi_i = 0, Q_i = 1) \\ = \; & P(\Delta_i = 1 | X_i, V_i, Q_i = 1) \\ = \; & \frac{\exp(\alpha' X_i + \gamma V_i)}{1 + \exp(\alpha' X_i + \gamma V_i)} I(Q_i = 1) \end{aligned} \tag{3.3}$$

That is, we estimate the probability of examiner-diagnosed TMD in a participant who was not examined as intended. (Here $I(x)$ denotes an indicator function.) The probability was estimated using the time between enrollment and their last positive screener as well as their answers on that screener. Then, for those individuals who screened positively on the last screener (i.e. those with $Q_i = 1$) and were not examined, the estimated probability of being a case is estimated by (3.3) with the parameters replaced by their respective estimates based on individuals who were examined.

Note that if there are repeated measures, we may use a generalized linear mixed effects logistic regression model rather than a standard univariate logistic regression model. For example, if some participants screen positively on more than one screener and are examined at least once, then we have multiple observations per participant. In that case, fitting a generalized linear mixed effects logistic regression model rather than a standard logistic regression model would account for correlations between the responses of the same participant.

3.3.3 Multiple Imputation

One popular method for handling missing data is multiple imputation. For a comprehensive review on multiple imputation, see (103). Our imputation procedure

is as follows:

(i) Estimate predicted probabilities as described in the previous section for observations with missing censoring indicators. These are the individuals who screen positively but are not examined by a clinician.

(ii) For each observation with a missing censoring indicator, generate a Bernoulli random variable with success probability equal to the predicted probability found in step (i).

(iii) Combine the raw data and imputed data from step (ii) to form a completed data set.

(iv) Fit the Cox proportional hazards model to the completed data set.

(v) Record each parameter estimate $\hat{\beta}_j$ and covariance matrix $\hat{U}_j$.

(vi) Repeat steps (ii)-(v) for a total of m times, where m is the desired number of imputations.

Next, we combine all of the estimates. The average parameter estimate is

$$\bar{\beta} = \frac{1}{m} \sum_{j=1}^{m} \hat{\beta}_j, \tag{3.4}$$

the within-imputation variance estimate is

$$\bar{U} = \frac{1}{m} \sum_{j=1}^{m} \hat{U}_j, \tag{3.5}$$

and the between-imputation variance

$$\hat{B} = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{\beta}_j - \bar{\beta})(\hat{\beta}_j - \bar{\beta})'. \tag{3.6}$$

Finally, the estimated covariance matrix is

$$\hat{Var}(\bar{\beta}) = \bar{U} + (1 + \frac{1}{m})\hat{B}. \tag{3.7}$$

3.3.4 Estimation of Incidence

Previous sections of this paper described how to estimate hazard ratios in the presence of missing censoring indicators. It may also be of interest to estimate incidence rates for the same event using Poisson regression instead of Cox regression. For example, one of the aims of the OPPERA study is to estimate the incidence rate of first-onset TMD.

In order to estimate incidence rates, we estimate the case probabilities as described previously based on participants who screened positively and were examined. Then we impute case status as described in section 3.3.3 for those who screened positively but were not examined. However, in this case we fit Poisson regression models, rather than Cox models, to the completed data sets. Finally, we calculate the incidence rate based on the estimates of the regression coefficients in the Poisson model. Specifically, we use the data from imputation j to fit the model

$$\log(E(\Delta_{ij}|Z_i, V_i)) = \alpha + \beta' Z_i + \log(V_i) \tag{3.8}$$

where Δ_{ij} denotes the j^{th} imputation for observation i, $j = 1, \ldots, m$. We combine the m imputations using equation (3.4) and

$$\bar{\alpha} = \frac{1}{m}\sum_{j=1}^{m} \hat{\alpha}_j. \tag{3.9}$$

The estimated incidence rate for an individual with covariate Z^* is given by $\exp(\bar{\alpha} + \bar{\beta} Z^*)$.

3.4 Simulations

Data with missing censoring indicators were simulated, and several possible methods were compared with respect to bias, coverage, and confidence interval width. Survival times for 1,000 individuals were generated with exponentially distributed failure times under a proportional hazards model with covariates as proposed by Bender et al. (4). That is, the survival time for each individual was distributed according to (3.1) where $\lambda_0(t) = 1$ is the baseline hazard. For our simulations, $Z_i = X_{i1}$ was a single baseline covariate following a normal distribution with mean 2 and unit variance. In other words, the failure times T_i followed an exponential distribution with hazard $\exp(\beta' X_{i1})$ where $\beta \in \{-0.5, -1.5, -3\}$. The censoring times C_i followed an exponential distribution with mean 5. This yielded about 35%, 75% and 90% censoring for $\beta = -0.5$, $\beta = -1.5$, and $\beta = -3$, respectively.

Covariates are represented by X_{i1}, a risk factor for TMD measured at enrollment, and X_{i2}, a measurement collected on the last screener. For each observation, a normally distributed covariate X_{i2} was generated with mean Δ_i and standard deviation 0.3, representing a continuous measure of the likelihood of being identified as an incident case of TMD. This was used to generate $Q_i = I(X_{i2} > 0.5)$, an indicator of whether participant i screened positive on their last screener. Also, $\xi_i = I(\Delta_i \text{ is observed})$ corresponds to the indicator of whether participant i came in for their clinical exam if $Q_i = 1$. In all simulations, we set $\Delta_i = 0$ if $Q_i = 0$ (i.e. the participant's final screener was negative). This decision was made to reflect the fact that OPPERA participants who screened negatively were not examined.

We created missing censoring indicators under the following classical missing data mechanisms of Rubin (101):

(I) The probability of having a missing censoring indicator is independent of the

data. This is known as missing completely at random (MCAR).

(II) The probability of having a missing censoring indicator depends on an observed covariate. This is known as missing at random (MAR).

(III) The probability of having a missing censoring indicator depends on the (potentially unobserved) censoring indicator. This is known as missing not at random (MNAR).

Our method assumes that the data are MAR, which includes MCAR as a special case. Our simulations under MAR and MNAR parallel the study protocol in that censoring indicators can only be missing for those with positive screeners. In other words, observations were potentially missing if and only if $Q_i = 1$. (Individuals with negative screeners have $Q_i = 0$ and are assumed to be censored. Those with positive screeners have $Q_i = 1$ and may have missing clinical examinations.) Details and results for MCAR and MNAR data are shown in Sections 6 and 6 in the Web Appendix. We also considered several simulation scenarios where the logistic regression model for predicting the censoring indicator was misspecified; see Section 6 in the Web Appendix. For MAR data, we set censoring indicators to be missing with probability

$$\begin{aligned} \rho_i(X_i, V_i) &= P(\xi_i = 0 | X_i, V_i, Q_i = 1) && (3.10) \\ &= \frac{\exp(-0.2 - 0.3X_{i1} + 0.1V_i)}{1 + \exp(-0.2 - 0.3X_{i1} + 0.1V_i)}. && (3.11) \end{aligned}$$

In each simulated data set, all observations with observed censoring indicators who had screened positively were used to fit a logistic regression model for case status with covariates X_{i1}, X_{i2} and V_i. That is, using the complete data (i.e. observations with $Q_i = 1$ and $\xi_i = 1$), we fit the logistic regression model for the event probability conditional on $X_i' = (1, X_{i1}, X_{i2})$ and V_i, namely

$$\text{logit}\{P(\Delta_i = 1 | X_i, V_i, Q_i = 1, \xi_i = 1)\} = \alpha' X_i + \gamma V_i \qquad (3.12)$$

The estimated probabilities $\hat{p}_i = \frac{\exp(\hat{\alpha}' X_i + \hat{\gamma} V_i)}{1+\exp(\hat{\alpha}' X_i + \hat{\gamma} V_i)}$ were calculated for individuals with $Q_i = 1$.

To evaluate the performance of our method, multiple imputation was employed to calculate 100 imputed estimates of β for each simulation as described in Section 3.3.3. For each observation i with $Q_i = 1$ and $\xi_i = 0$, we estimated failure indicators $\hat{\Delta}_{ij} \sim \text{Bernoulli}(\hat{p}_i)$ independently for each imputation j.

A Cox proportional hazards model was fit for each imputed data set, and the imputed estimates of the regression coefficient and their variances were recorded. These were aggregated using equations (3.4) and (3.7) to create confidence intervals for the multiple imputation estimates.

The performance of our method was compared with that of the method of Cook and Kosorok (8). To obtain the estimates of (8), for each simulated data set, we estimated the probabilities $\hat{p}_i$ that the (potentially unobserved) event for participant i is a true event, as described previously. We then fit a weighted Cox proportional hazards model to the data set. Two new observations were generated for observations with missing censoring indicators. Each such pair of observations had the same failure time and covariates, but different failure indicators and weights. The first observation had weight $\hat{p}_i$ and $\hat{\Delta}_i = 1$, and the second observation had weight $1 - \hat{p}_i$ and $\hat{\Delta}_i = 0$. Participants with fully observed data had unit weight. The estimated regression coefficient, $\hat{\beta}$ was recorded. The variance of this estimate was estimated by generating 1000 bootstrap replicates of each simulated data set and refitting the model for each bootstrap replicate. The average parameter estimate, $\bar{\hat{\beta}}$ and percentile confidence intervals $(\beta_{0.025}, \beta_{0.975})$ were all recorded, where β_α is the α^{th} quantile among the 1000 bootstrap replicates.

We also compared our method to the ideal situation in which all data were observed, complete case analysis (meaning that we exclude from the data set all observations with missing censoring indicators), and two ad hoc methods in which we treat the missing

indicators either all as censored or all as failures. Results under the assumption of MAR are shown in Table 3.1. We estimated the bias of each method by calculating the mean difference between the estimated Cox regression coefficient and the true coefficient over the 1000 simulations. We also calculated the mean width of the confidence intervals produced by each method over the 1000 simulations. Similarly, we calculated the empirical coverage probability for the confidence intervals produced by each method by dividing the number of times that the confidence intervals contained the true value of the parameter by 1000. Finally, we report the Monte Carlo error for the coverage rate, which is the error in the empirical coverage probability due to conducting only a finite number of simulations (which would be $\sqrt{\alpha(1-\alpha)/n}$ for n simulations).

The empirical coverage probability of the imputed confidence intervals is close to the nominal level (0.95) in all simulations. Our multiple imputation method and the method of Cook and Kosorok (8) produced approximately unbiased estimates and valid confidence intervals in all the scenarios we considered. The estimates produced by the other methods showed a larger amount of bias and did not always achieve the desired coverage level. Our multiple imputation method also yielded the narrowest confidence intervals in each scenario, although the method of Cook and Kosorok (8) produced confidence intervals that were only slightly wider. Moreover, for most parameter values, the coverage probabilities for the complete case and ad hoc methods were significantly different ($p < 0.01$) from the nominal rate.

In addition, we examined the performance of our proposed methods when we changed the logistic regression model for Δ_i. We investigate two additional types of models: one in which the model contained a variable unrelated to case status and another in which the model does not include one variable related to case status. As in the previous simulations, the failure times were generated by (3.1), censoring was exponential with mean 5, failure indicators were set to be missing completely at

random or missing at random with probability given in equation (3.10), $X_{i1} \sim N(2,1)$, $X_{i2} \sim N(\Delta_i, 0.3)$ and $Q_i = I(X_{i2} > 0.5)$ for $i = 1, \ldots, n$. We also generated $X_{i3} \sim N(0,1)$ where X_{i1}, X_{i2}, X_{i3} were mutually independent.

In the previous simulations, we fit the data to (3.12) with $X_i = \{X_{i1}, X_{i2}\}$. The additional simulations instead used the covariates and parameters as follows:

(A) $\tilde{X}_i = \{1, X_{i1}, X_{i2}, X_{i3}\}$, $\tilde{\alpha} = \{\tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3\}$

(B) $\tilde{X}_i = \{1, X_{i1}\}$ $\tilde{\alpha} = \{\tilde{\alpha}_0, \tilde{\alpha}_1\}$.

That is, rather than fitting model (3.12) to the data, we modeled the case probability with

$$\text{logit}\{P(\Delta_i = 1 | X_i, V_i, Q_i = 1)\} = \tilde{\alpha}' \tilde{X}_i + \gamma V_i. \tag{3.13}$$

The results, which are shown in Section 6 in the Web Appendix, remained similar under both alternative models. This indicates that the proposed methods are robust to misspecification of the logistic regression model. Most notably, leaving out one covariate that was weakly related to case status did not markedly decrease the performance of the method.

Finally, we conducted simulations to evaluate the method's ability to estimate incidence rates. A similar multiple imputation strategy was applied to Poisson regression. Our method produced estimates much closer to the true incidence rates than the complete case estimate. In fact, the complete case method underestimated incidence rates by as much as a factor of 3. See Section 6 in the Web Appendix for details.

3.5 Analysis of the OPPERA Study

In this section, we apply our method to estimate hazard ratios and incidence rates in the OPPERA study.

3.5.1 Hazard Ratios

We applied our method to the OPPERA cohort to adjust for the effect of participants with missing clinical examinations. (Note that examinations for participants evaluated by Examiner #4 were also treated as missing.) First, we estimated the probability that a participant would be diagnosed as an incident case of TMD given a positive screener. Due to the rich body of information collected in each screener, we carefully selected a small number of predictor variables. Specifically, we fit a generalized linear mixed model with a logistic link function to predict the result of the clinical exam based on each item in the screener. A mixed model was used because a significant (n=113) minority of participants screened positive more than once. All models were adjusted for study site and included a random effect term for each participant.

The majority of the variables measured on the screener were not associated with the result of the clinical examination. The strongest predictor of being diagnosed with TMD was a count of non-specific orofacial symptoms (e.g stiffness, fatigue) in the previous three months. The time elapsed since enrollment and OPPERA study site were also important covariates, as shown in (1). Several other possible predictors of being diagnosed with TMD were identified, but including these additional predictors in the model did not significantly improve the predictive accuracy of the model. Thus, we estimated the probability of being diagnosed with TMD based on the count of non-specific orofacial symptoms, time since enrollment, and OPPERA study site. This model was used to perform multiple imputation for those with no clinical examination. These imputed data sets were used to fit a series of Cox proportional hazards models to estimate the hazard ratio (and associated confidence interval and p-value) for each predictor using the methods described in section 3.3.3. Examples of predictors include perceived stress, history of comorbid chronic pain conditions, and smoking status.

In addition, Bair et al. (1) examined univariate relationships between examination

attendance and numerous possible predictor variables. Differences between examined and non-examined participants were small and most were not statistically significant. However, this indicates that the data were not missing completely at random.

Table 3.2 shows the results of applying our method to a subset of the putative risk factors of TMD measured in OPPERA. Due to the large number of putative risk factors measured in OPPERA, we only report the results for a selected subset of the variables. All continuous variables were normalized to have mean 0 and standard deviation 1 prior to fitting the Cox models. (Thus, the hazard ratios for the continuous variables represent the hazard ratios corresponding to a one-standard deviation increase in the predictor variable.) In Table 3.2, all the quantitative sensory testing and psychosocial variables were continuous, while all of the clinical variables were dichotomous (and hence were not normalized). For a more detailed description of the OPPERA domains, see Section 6 in the Web Appendix, (75), and (113).

Compared to the unimputed results, which treated missing censoring indicators as censored observations, imputation slightly reduced the hazard ratios for most of the psychosocial variables that were measured in OPPERA. For instance, Table 3.2 shows the (standardized) hazard ratios for the Pennebaker Inventory of Limbic Languidness (PILL) score, the neuroticism subscale of the Eyesenk Personality Questionnaire (EPQ), the Spielberger Trait Anxiety Inventory score, the Perceived Stress Scale, and the somatization subscale of the Symptom Checklist-90, Revised (SCL-90R). In each case, the hazard ratios were reduced after imputation.

A similar pattern was observed after applying our imputation method to the measures of experimental pain sensitivity. The mechanical pain aftersensation ratings were strongly associated with first-onset TMD before imputation, but they were only weakly associated with first-onset TMD after imputation. The pressure pain algometer ratings were also more weakly associated with TMD after imputation (and two of three ratings

in Table 3.2 were no longer significantly associated with first-onset TMD at the $p < 0.05$ level).

Interestingly, the hazard ratios for the presence of one or more palpation tender points at the temporalis and masseter muscles were also attenuated after imputation. These tender points were evaluated as part of the clinical examination using a different protocol than the quantitative sensory testing algometer pain ratings. However, both pain measures (algometer and palpation) were measured at the same facial locations. While the palpation ratings were more strongly associated with first-onset TMD than the algometer ratings both before and after imputation, it is interesting that different pain sensitivity measures using different protocols at the same anatomical location were both attenuated by imputation.

The effects of other clinical variables were also attenuated after imputation. For example, the hazard ratios associated with being unable to open one's mouth wide in the past month and having two or more comorbid pain conditions were both noticeably attenuated after imputation. However, other clinical variables were more strongly associated with first-onset TMD after imputation. For example, having a history of respiratory illness was only weakly associated with first-onset TMD before imputation (HR=1.38, p=0.04), but the association was much stronger after imputation (HR=1.43, p=0.004). Also, being a current smoker was not significantly associated with first-onset TMD before imputation (HR=1.26, p=0.24) but was associated after imputation (HR=1.49, p=0.02).

3.5.2 Incidence Rates

In Table 3.3, the incidence rate of first-onset TMD was estimated using two different approaches. First, all missing censoring indicators were treated as censored. Second, the multiple imputation method in this paper was used to estimate the incidence rate.

The estimated TMD incidence rate using multiple imputation was 66% greater than the unimputed estimate. The estimated incidence rate increased by 64% for females and 72% for males. Estimated incidence rates for whites and Hispanics were 99% and 193% higher, respectively, with imputation. Thus, the incidence rate is likely to be underestimated without imputation.

3.6 Discussion

We have developed a computationally efficient method to adjust for missing censoring indicators in time-to-event data using logistic regression and multiple imputation. Logistic regression is used to estimate the failure probability for participants with missing censoring indicators. The missing values are imputed, and the standard errors are estimated using our multiple imputation method. This framework is important in studies where failure status may be measured in stages, which may lead to missing censoring status indicators. This is a common occurrence in studies of diseases that are difficult or expensive to diagnose, such as TMD.

The present method is similar to the method of Magder and Hughes (74), who use an iterative procedure for parameter estimation based on the EM algorithm. Our assumption of MAR data renders their iterative method unnecessary. Other methods (78, 43, 122) depend on the MCAR assumption, which does not hold for the OPPERA study. Chen et al. (6) estimate Cox regression parameters using the EM algorithm and establish their consistency under basic regularity conditions, including missing at random (MAR) censoring indicators. However, their approach depends on the assumptions of piecewise constant proportional hazard functions for the censoring time as well as for the failure time.

In each simulation scenario, our multiple imputation method produced the narrowest

valid confidence intervals and no significant bias. In particular, the method of Cook and Kosorok (8) produced slightly wider confidence intervals in all but one of the simulations we considered. The differences were extremely small, so the performance of the two methods appear to be comparable for most practical purposes. However, we believe that our method has several possible advantages over the method of Cook and Kosorok (8). First, bootstrapping is much more intensive computationally than our multiple imputation approach. Calculating bootstrap confidence intervals generally requires at least 1000 bootstrap replicates (33), whereas as few as 10 imputed data sets may be sufficient for multiple imputation (70). Although the difference in the computing time of the two methods is small for a single fitted model, many such models will be required in the course of the OPPERA study. OPPERA has already collected data on approximately 3000 genetic markers and has plans to collect data on approximately a million genetic markers in a genome-wide association study. Thus, at least 3000 (and potentially as many as a million) Cox models will need to be fit, and our proposed method may allow for a significant decrease in computing time. Moreover, our method can also be easily implemented in popular statistical software packages (such as SAS) without additional programming.

Additionally, our methodology may easily be extended to other models, such as Poisson regression. We conducted simulations (Table A1.10 in the Web Appendix) that showed that our proposed method can be used to estimate incidence rates using Poisson regression, which is one of the research aims of the OPPERA study. In particular, estimates of the failure rates were biased when missing censoring indicators were treated as censored or when the complete case method was used, but they were unbiased when we employed the methodology in this paper. The method of Cook and Kosorok (8) cannot be used for incidence rate estimation.

Our method may yield increased bias and decreased coverage if the logistic regression

model for predicting case status is inaccurate, as observed in the simulations in Section 6 in the Web Appendix. However, this would also be true for competing methods, including the method of (8).

In the OPPERA study, the hazard ratios associated with some variables were noticeably different after imputation. Although other results remained qualitatively unchanged, we note that even small changes in hazard ratios are important. In addition, estimated incidence rates were significantly increased after imputation. Since the results of OPPERA may become normative in the orofacial pain literature, precise calculation of the incidence rate of TMD and the hazard ratios associated with putative risk factors is important. Thus, imputation is recommended.

Table 3.1: Simulation Results for MAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0005	0.0006	0.1668	0.0001	0.938
	Complete Case	0.0018	0.0007	0.2155	0.0001	0.951
	Treat all as Censored	0.1053	0.0007	0.2131	0.0001	0.494
	Treat all as Failures	0.0018	0.0006	0.1701	0.0001	0.941
	Cook & Kosorok	-0.0010	0.0006	0.1738	0.0001	0.943
	Multiple Imputation	-0.0010	0.0006	0.1716	0.0001	0.938
-1.5	Full Data	-0.0008	0.0010	0.3185	0.0002	0.966
	Complete Case	-0.0626	0.0015	0.4343	0.0003	0.930
	Treat all as Censored	0.1215	0.0014	0.4229	0.0003	0.778
	Treat all as Failures	0.0680	0.0010	0.3160	0.0002	0.852
	Cook & Kosorok	0.0002	0.0011	0.3412	0.0004	0.951
	Multiple Imputation	0.0002	0.0011	0.3309	0.0002	0.961
-3	Full Data	-0.0301	0.0025	0.7627	0.0009	0.957
	Complete Case	-0.1996	0.0037	1.0840	0.0017	0.913
	Treat all as Censored	0.0987	0.0035	1.0417	0.0016	0.919
	Treat all as Failures	0.5875	0.0024	0.6307	0.0006	0.104
	Cook & Kosorok	-0.0275	0.0027	0.9112	0.0017	0.946
	Multiple Imputation	-0.0282	0.0027	0.8057	0.0011	0.947

*: The Monte Carlo error is 0.007.

Table 3.2: Results from the OPPERA Study

	Treat All MCIs as Censored				**Multiple Imputation**			
	HR	LCL	UCL	P	HR	LCL	UCL	P
Clinical Variable								
In the last month could not open mouth wide	3.26	1.83	5.84	<0.0001	2.45	1.42	4.22	0.0012
Has two or more comorbid chronic pain disorders	3.08	2.26	4.21	<0.0001	2.50	1.90	3.29	<0.0001
History of 5 respiratory conditions	1.38	1.01	1.87	0.0408	1.45	1.13	1.87	0.0040
Smoking: current	1.26	0.86	1.84	0.2403	1.49	1.07	2.09	0.0199
Smoking: former	1.87	1.22	2.87	0.0041	1.65	1.12	2.43	0.0106
One or more palpation tender points: right temporalis	1.83	1.32	2.52	0.0002	1.54	1.18	2.02	0.0017
One or more palpation tender points: left temporalis	1.60	1.14	2.25	0.0064	1.48	1.12	1.97	0.0060
One or more palpation tender points: right masseter	1.85	1.35	2.53	0.0001	1.63	1.25	2.12	0.0003
One or more palpation tender points: left masseter	1.70	1.23	2.35	0.0013	1.53	1.17	2.01	0.0021
Quantitative Sensory Testing Variable								
Pressure pain threshold: temporalis	1.26	1.07	1.49	0.0065	1.16	1.01	1.33	0.0335
Pressure pain threshold: masseter	1.23	1.04	1.45	0.0170	1.15	1.00	1.32	0.0576
Pressure pain threshold: TM joint	1.25	1.05	1.48	0.0106	1.14	1.00	1.30	0.0555
Mechanical pain aftersensation: 512mN probe, 15 s	1.23	1.09	1.38	0.0006	1.15	1.03	1.28	0.0123
Mechanical pain aftersensation: 512mN probe, 30 s	1.20	1.07	1.34	0.0020	1.12	1.01	1.25	0.0328
Psychosocial Variable								
PILL Global Score	1.52	1.35	1.71	<0.0001	1.46	1.31	1.62	<0.0001
EPQ-R Neuroticism	1.39	1.21	1.60	<0.0001	1.26	1.12	1.42	0.0002
Trait Anxiety Inventory	1.43	1.25	1.64	<0.0001	1.35	1.21	1.52	<0.0001
Perceived Stress Scale	1.35	1.17	1.55	<0.0001	1.30	1.16	1.47	<0.0001
SCL 90R Somatization	1.44	1.31	1.58	<0.0001	1.40	1.29	1.52	<0.0001

Table 3.3: Estimated TMD Incidence Rates With and Without Imputation

	No MI	MI	Percent Change
Overall	2.23	3.70	66
Males	1.87	3.22	72
Females	2.46	4.03	64
White	1.70	3.37	99
Black	4.20	5.32	27
Hispanic	1.17	3.44	193
Other	1.10	1.80	63

Incidence rates are given in cases per 100 person-years.

CHAPTER 4: MODELING SECONDARY PHENOTYPES CONDITIONAL ON GENOTYPES IN CASE-CONTROL STUDIES

4.1 Introduction

Prospective studies are more straightforward and less prone to confounding than other study designs. However, they may require either extremely long follow-up periods or large sample sizes and lack power. For rare diseases, in particular, the sample sizes required in a prospective cohort study to have adequate statistical power to test hypotheses of interest may be prohibitively large. This can be especially problematic in genetic association studies, which may cost thousands of dollars per participant just to extract their genetic profiles. Retrospective case-control studies are more cost-effective. The number of case-control studies focusing on the relationship between genetics and disease outcomes has grown astronomically in recent years.

It is well known that when modeling the probability of case status in a case control design, simple logistic regression may be used to model the primary outcome as if the study were prospective (95). However, researchers may design studies based on one outcome but study outcomes of secondary interest simultaneously or via a new follow-up study. Without proper care, analysis of secondary phenotypes in case-control studies may be problematic. Standard methods, such as logistic regression, may be biased, inefficient, or lead to misleading inference. The standard method of unweighted regression on the full case control sample and the method of adjusting for case status with an indicator variable have inflated type I error when the disease is not rare, (82).

The popular practices of restricting to cases or restricting to controls greatly reduce efficiency and may be subject to bias.

This methodological work arose in consideration with data from the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) Study (113). The OPPERA study was primarily designed to identify risk factors for temporomandibular disorders (TMD). In addition to the cohort of initially-TMD-free adults enrolled in the prospective cohort study, people with examiner-verified chronic TMD were were enrolled to create an unmatched, case-control study. A large number of putative risk factors were collected at enrollment (75). In particular, investigators seek to explain relationships between TMD and other chronic pain conditions. One putative risk factor of interest in its own right is the (ordinal) number of comorbid pain conditions a subject experiences. The genetic information collected may be predictive of comorbid conditions as well as of TMD. There currently are no readily available methods to assess the relationship between comorbid conditions, a secondary phenotype, and individual candidate genes.

Over the past five years, some methods have been proposed for analyzing secondary phenotypes in case-control studies. For logistic regression, (97) recommends weighting subjects from a nested prospective case-control study by the reciprocal of their probability of selection. This stratum-weighted logstic regression method, also called inverse probability weighting (IPW), achieves the nominal type I error rate but can be less efficient than the standard unadjusted method or the method of adjusting for case status (82). (Yet, in light of the fact that the method of adjusting for case status may have inflated type I error, the lower power of IPW is less alarming.) More significantly, the IPW estimator of (97) may merit a correction factor for the standard error. (82) provides a robust sandwich estimator for the variance based on generalized estimating equations (GEE), which applies to both continuous and binary outcomes.

(66) propose a likelihood based method for both continuous and binary outcomes that is more powerful than the IPW method, but the results may be biased and have inflated size when there is significant interaction between the genotypes and the original outcome (63, 47). (140) propose a bootstrap estimate for binary outcomes. They first derive non-linear equations relating the logistic regression coefficients to known sample sizes and prevalence estimates for the primary disease and secondary phenotype. Then they calculate the unadjusted parameter estimate and standard error, resample parameter estimates from a normal distribution with that mean and standard error, refine the estimate for each replication by solving the aforementioned non-linear equations, and then using the percentile method for confidence intervals. They also adapt their bias correction method to the frequency-matched case-control study design and extend the IPW method to retrospective case-control studies using an estimate of primary disease prevalence to calculate weights. (141) demonstrate that these methods have greater efficiency than the IPW method, but they assume that the secondary phenotype and covariates are binary and there is no gene-environment interaction.

When there is gene-environment interaction and the disease is rare (63) recommend estimation with adaptive weighting motivated by a Bayesian shrinkage estimator. (47) discuss joint modeling based on Gaussian copulas for the analysis of multiple secondary phenotypes in the exponential family. The copula method has controlled type I error and is more powerful than the IPW method.

The aforementioned methods may not be appropriate for all applications. First, the IPW method of (97) may require an adjustment to the standard error and was originally only designed to apply to binary outcomes. As far as we know, there are no methods that can be applied to secondary phenotypes outside the exponential family. Time-to-event outcomes, for example, may be of clinical interest, and may be present in large studies, such as OPPERA. Sequencing studies also utilize more complicated

test statistics outside of the exponential family. Additionally, there is a lack of software that may be easily implemented.

In this paper, we propose a method for analyzing secondary phenotypes in case-control genetic association studies. We advocate using the IPW method of (97) for parameter estimation, but estimating the standard error via bootstrapping. This maintains the simplicity and intuitiveness of IPW and generalizes it to a wider variety of situations than previously considered, while providing a valid estimate of the standard error. Our method can handle arbitrary types of analyses, including time-to-event and nonparametric methods, as well as logistic regression and linear models as described in the current literature. Moreover, unlike other methods in the literature, IPW can be easily generalized to outcomes for which no existing method applies. We describe our methodology in detail in section 4.2. Simulations are presented in section 4.3. The method is applied to the OPPERA study in section 4.4. We conclude with a discussion.

4.2 Proposed Method

Consider a case-control study consisting of n cases and m controls. Let Z_i, a $p \times 1$ vector, denote covariate information, D_i denote the case-control status (1=case, 0=control), and Y_i denote the secondary phenotype for $i = 1, \ldots, n+m$. In the OPPERA study, Z_i denotes the number of copies of the minor allele, D_i is an indicator of whether participant i is a chronic case of TMD, and Y_i is the number (0,1,2+) of comorbid pain conditions for participant i. If one were to ignore the case-control study design and consider the data as a random sample from the population, then one would use standard methodology to study the relationship between $Y_i = (Y_1, \ldots, Y_{n+m})'$ and $Z = (Z_1, \ldots, Z_{n+m})'$. Denote the log-likelihood under the assumption of random sampling as $l(\theta|Y_i, Z_i)$ where θ is a $q \times 1$ vector of model parameters.

The IPW method of (97) simply weights standard analyses appropriately to account

for over-sampling of cases. Specifically, in a prospective (nested) case-control study, if we denote f_{ca} as the sampling fraction for cases and f_{co} as the sampling fraction for controls, we use $w_i = 1$ as the weight for cases and $w_i = \frac{f_{ca}}{f_{co}}$ as the weight for controls. For retrospective case-control studies, the weights may be estimated as in (140) by $w_i = 1$ for cases and $w_i = \frac{n(1-p_e)}{mp_e}$ for controls. We may write $w_i(D_i) = D_i + (1 - D_i)w_i$.

The weighted log-likelihood is the weighted sum of the log-likelihood of each observation

$$l_W(\theta|Y, Z, D) = \sum_{i=1}^{m+n} w_i(D_i) l(\theta|Y_i, Z_i) \tag{4.1}$$

where $l(\theta|Y_i, Z_i) = log[f(Y_i|Z_i)]$ if Y is continuous or $l(\theta|Y_i, Z_i) = log[P(Y_i|Z_i)]$ if Y is discrete. For example, if Y_i is binary, one would typically use weighted logistic regression with

$$l_W(\theta|Y, Z, D) = \sum_{i=1}^{m+n} log[P(Y_i = 1|Z_i, D_i)] = \sum_{i=1}^{m+n} w_i(D_i)\{Y_i \alpha' Z_i - log[1 + exp(\alpha' Z_i)]\} \tag{4.2}$$

If Y_i is continuous, a weighted linear model could be utilized,

$$l_W(\theta|Y, Z, D) = \sum_{i=1}^{m+n} log[f(Y_i|Z_i, D_i)] = \sum_{i=1}^{m+n} w_i(D_i)[-\frac{log(2\pi\sigma^2)}{2} - \frac{(Y_i - \alpha' Z_i)^2}{2\sigma^2}] \tag{4.3}$$

where $Var(Y_i) = \epsilon_i \sim N(0, \sigma^2)$.

If (Y_i, Δ_i) is a (possibly censored) time-to-event outcome with failure time T_i and censoring time C_i and $Y_i = min(T_i, C_i)$ and $\Delta_i = I(T_i < C_i)$, then one may use a proportional hazards model,

$$\lambda_{T_i}(t|Z_i) = \lambda_0(t) exp(\beta' Z_i) \tag{4.4}$$

with weighted log-partial-likelihood given by

$$l_W(Y|Z,D) = \sum_{i=1}^{m+n} w_i(D_i)\Delta_i\{\beta' Z_i - log[\sum_{l=1}^{m+n} w_l(D_l)I(Y_i < Y_l)exp(\beta' Z_l)]\} \qquad (4.5)$$

We propose the use of bootstrapping to estimate the standard error of the estimate of interest. We select R samples from the empirical distribution of the original data. For each bootstrap replication, we apply the IPW method described above.

Specifically, let (D, Y, Z) denote the data with empirical cdf f, and let (D_r^*, Y_r^*, Z_r^*) denote bootstrap replications for $r = 1, \ldots, R$. The first step is to fit a model to (D_r^*, Y_r^*, Z_r^*) using the weighted log likelihood (4.1) for each replication. The variance of the parameter estimate is given by the estimated variance of the R bootstrap parameter estimates. Confidence intervals may be generated by the percentile method, BCa method, bootstrap-t, or a normal approximation. Standard software, such as the boot package in R, can easily generate these estimates.

4.3 Simulation Study

4.3.1 General Setup

We simulated data under the framework in (66). In order to have n cases and m controls, we generated $n_s = \frac{3max(m,n)}{p_e}$ observations where p_e is the estimated prevalence of cases in the population. This ensured there would be enough cases and controls in each dataset. Next, we generated case status as detailed below. Lastly, we used rejection sampling to select a subset of the cases and controls of the desired sample size.

For all subjects in the large dataset, $i = 1, \ldots, n_s$, we assume that the relationship between case-control status, D_i, the number of copies of the minor allele, Z_i, and

the secondary phenotype, Y_i, is given by a logistic regression model based on the genetic profile and secondary phenotype. The distribution of each type of outcome and corresponding specific form of the logistic regression model are given in sections 4.3.2, 4.3.3, and 4.3.4 for binary, ordinal, and time-to-event outcomes, respectively. Each equation specifies that the probability of being a case of the primary disease (rather than a control) depends on the status of the genetic profile and the secondary phenotype. The inheritance model is assumed to be additive. We assumed an estimated prevalence of D of 10% and a minor allele frequency of 30%.

Finally, we form a dataset consisting of the first n cases and m controls. For each of these simulations, we generated 1000 datasets each with $n = 1000$ cases and $m = 1000$ controls. For each method and scenario, we estimated the value of the parameter relating the secondary phenotype and the number of copies of the minor allele. (This was the log-odds ratio or log-hazard ratio depending on the type of outcome.) Average bias, empirical coverage, and average confidence interval width were compared between our method, and the naive methods that restrict to cases, restrict to controls, or adjust for case-control status using an indicator variable. We also compared performance to the IPW with GEE method of (82) when applicable, i.e. for continuous outcomes.

4.3.2 Continuous Phenotypes

For continuous secondary phenotypes, we assume a standard linear model with normally distributed errors, $\epsilon_i \sim N(0, \sigma^2)$,

$$Y_i = \beta_0 + \beta_1 Z_i + \epsilon_i. \tag{4.6}$$

where X_i is defined in section 4.3.1, and case status is defined by

$$logit[P(D_i = 1|Y_i, Z_i)] = \gamma_0 + \gamma_1 Z_i + \gamma_2 Y_i. \tag{4.7}$$

Our simulations included the parameters $\beta_0 = 0$, $\beta_1 = -0.12, -0.5, -1, -2$ and $\gamma_1 = log(2), log(3), log(5), log(10)$, $\gamma_2 = log(2)$ and $\sigma^2 = 1$.

In order to keep the prevalence approximately constant, we set the value of γ_0 separately for each simulation, according to

$$\hat{\gamma_0} = log(\frac{p_e exp(-1 * \tilde{X})}{1 - p_e}) \tag{4.8}$$

where

$$\tilde{X} = \gamma_1 \bar{Z} + \gamma_2 \bar{Y}. \tag{4.9}$$

and $\bar{Z} = \frac{1}{n_s} \sum_{i=1}^{n_s} Z_i$ and $\bar{Y} = \frac{1}{n_s} \sum_{i=1}^{n_s} Y_i$ are the averages of the $i = 1, \ldots, n_s$ Z_i and Y_i values.

The parameter of interest was β_1. Considering continuous outcomes facilitated comparison to the IPW with GEE method of (82).

Simulations with continuous outcomes yielded the following results. In all scenarios, our method had negligible bias and coverage rate near 95%, as desired. Performance in terms of bias, coverage, and confidence interval width was comparable to that of (82). The bootstrapping-IPW method had comparable bias to the method of (82) and less bias than all other methods. Details are found in Table 4.1. This shows when competing method are applicable, our method does at least as well as, if not better than the competitors.

4.3.3 Ordinal Phenotypes

We tested four scenarios for ordinal phenotypes with 3 levels. For simplicity, we will denote these $Y_i = 0$, $Y_i = 1$ and $Y_i = 2$. In general, we generated the ordinal outcomes

Table 4.1: Results of Simulations for Intermediate Continuous Phenotypes

Parameters	β_1 = -0.12 γ_1 = log(2)	β_1 = -0.12 γ_1 = log(3)	β_1 = -0.12 γ_1 = log(5)	β_1 = -0.12 γ_1 = log(10)	β_1 = -0.5 γ_1 = log(2)	β_1 = -1 γ_1 = log(2)	β_1 = -2 γ_1 = log(2)
LM (Bias)	-0.044	0.059	0.061	0.048	0.028	0	-0.058
LM, controls only (Bias)	-0.039	-0.073	-0.108	-0.151	-0.022	-0.002	0.037
LM, cases only (Bias)	-0.056	-0.099	-0.166	-0.256	-0.028	0.002	0.061
LM adjusted for case status (Bias)	0.048	-0.088	-0.14	-0.208	-0.025	0	0.047
Monsees (Bias)	-0.002	-0.001	0.001	0.001	0.001	-0.001	-0.003
Bootstrap (Bias)	-0.002	-0.001	0.001	0.001	0.001	-0.001	-0.003
LM (Coverage)	0.728	0.577	0.532	0.666	0.883	0.944	0.643
LM, controls only (Coverage)	0.875	0.698	0.444	0.206	0.926	0.939	0.892
LM, cases only (Coverage)	0.761	0.394	0.06	0	0.911	0.949	0.794
LM adjusted for case status (Coverage)	0.696	0.235	0.021	0	0.877	0.945	0.736
Monsees (Coverage)	0.950	0.951	0.948	0.949	0.952	0.938	0.949
Bootstrap (Coverage)	0.950	0.956	0.944	0.946	0.95	0.937	0.944
Monsees (Width)	0.164	0.160	0.155	0.148	0.166	0.167	0.168
Bootstrap (Width)	0.162	0.160	0.154	0.147	0.165	0.167	0.168

LM=Linear Model

with the following probabilities

$$p_0 = P(Y_i = 0) = exp(\zeta_0 + \beta Z_i)/(1 + exp(\zeta_0 + \beta Z_i))$$

$$p_1 = P(Y_i = 1) = exp(\zeta_1 + \beta Z_i)/(1 + exp(\zeta_1 + \beta Z_i)) - p_0$$

$$p_2 = P(Y_i = 2) = 1 - p_1 - p_0$$

For all subjects, $i = 1, \ldots, m + n$, we assume that the relationship between case-control status, D_i, the number of minor allele copies, Z_i, and the secondary phenotype, Y_i, is given by the following logistic regression model

$$logit[P(D_i = 1|Y_i, Z_i)] = \gamma_0 + \gamma_1 Z_i + \gamma_{2a} I(Y_i = 1) + \gamma_{2b} I(Y_i = 2). \tag{4.10}$$

where

$$\bar{Y} = \gamma_{2a}(\frac{\sum_{i=1}^{m+n} I(Y_i = 1)}{m + n}) + \gamma_{2b}(\frac{\sum_{i=1}^{m+n} I(Y_i = 2)}{m + n}) \tag{4.11}$$

is used to define the average outcome in equation (4.9) and thus the value of γ_0 in (4.8).

For the four scenarios, we used the following parameters:

1. $\beta = 0.5$, $\zeta_0 = 1.5$, $\zeta_1 = 2.5$, and $\gamma_1 = \gamma_{2a} = \gamma_{2b} = log(2)$

2. $\beta = 1$, $\zeta_0 = 0$, $\zeta_1 = 1$, and $\gamma_1 = \gamma_{2a} = \gamma_{2b} = log(2)$

3. $\beta = 0.75$, $\zeta_0 = 1$, $\zeta_1 = 2$, and $\gamma_1 = \gamma_{2a} = \gamma_{2b} = log(2)$

4. $\beta = 0.5$, $\zeta_0 = 1.5$, $\zeta_1 = 2.5$, $\gamma_1 = \gamma_{2a} = log(2)$, and $\gamma_{2b} = log(3)$.

Our weighted bootstrap method has less bias than all other methods. None of the other methods have adequate coverage for these ordinal simulations. Results are given in Table 4.2.

4.3.4 Time-to-Event Phenotypes

Survival outcomes were generated as in (4) with exponential failure and censoring times. The failure time T_i satisfies equation (4.4) where $\lambda_0(t) = 1$ for all t, $\beta = -1$ and Z_i is the minor allele frequency. The censoring time was exponential with shape parameter 2. The parameter of interest was β and the outcome of interest was (Y_i, Δ_i) where $Y_i = min(T_i, C_i)$ and $\Delta_i = I(T_i < C_i)$. This yielded about 84% censoring.

Case status in similar to equation (4.7) for continuous outcomes, but instead depended on the true failure time rather than the observed time as follows

$$logit[P(D_i = 1|X_i, T_i)] = \gamma_0 + \gamma_1 X_i + \gamma_2 T_i. \tag{4.12}$$

The value of γ_0 was set by equation (4.8) with $\tilde{X}$ defined by equation (4.9) and $\bar{X}$ and $\bar{Y}$ defined as in section 4.3.2. We used $\gamma_1 = \gamma_2 = log(2)$.

For time-to-event outcomes, our method retained empirical coverage around 95% and had less bias than all other methods. None of the other methods have adequate coverage, except the method that adjusts for case status. However, the latter method was overly conservative. See Table 4.3 for details. Other methods do not apply for this type of outcome. Consequently, no comparison is made.

4.4 Data Application

We applied the method to baseline case-control genetic study within OPPERA. The prospective cohort study consisted of 3263 healthy TMD-free volunteers and 186 volunteers determined at baseline to have TMD. All 186 cases were retained and 1633 controls were randomly selected into the baseline case-control study.

The covariates of interest were 2924 SNPs collected in a genetic association study of 3037 participants (115). The outcome was the number of co-morbid conditions,

categorized as either zero, one, or more than one co-morbid condition. Upon enrollment in OPPERA, participants self-reported by checking experience with a list of 20 conditions on the Comprehensive Pain and Symptom Questionnaire (CPSQ). Examples of chronic pain conditions include arthritis, fibromyalgia, irritable bowel syndrome, include chronic pelvic pain, among others. All cases and 1626 controls filled out the CPSQ, (87). Combining these yielded 166 cases and 1435 controls with information available on both their history of comorbid conditions and their genetic profiles. Recruited from 4 study sites and ranged from 18 to 44 years in age, these 1601 individuals comprise the proceeding analysis. For more details on the OPPERA study design, see (113, 75, 115, 87).

For each SNP with less than 5% missing values, we fit a proportional odds model to the data, adjusting for study site, age, gender, and two racial eigenvectors.

We collected the p-values and created QQ-plots of the negative logarithm of the p-values for the standard unweighted method and for our weighted bootstrapping method. The plots indicate that neither method found any SNPs that were significantly associated with comorbid pain conditions after adjusting for multiple comparisons. See Figures 4.1 and 4.2.

4.5 Discussion

Our proposed method for analysis of intermediate phenotypes in case-control studies of genetic data is simple and easily implemented in standard software. The simulation results indicate that it is approximately unbiased, has comparable coverage and confidence interval width to the method of IPW with GEE by (82). Under situations in which the method of (66) is applicable, their method should be more powerful than our proposed method.

Our method is general enough to allow for analysis of multiple outcomes simultaneously

and for outcomes for which previous methodology not applicable. Multiple outcomes could be analyzed using standard multivariate methods but weighting each observation as described in this paper, and bootstrapping to estimate the standard error. Currently, our method is the only viable way to evaluate secondary time-to-event outcomes in case-control studies. More importantly, the method can be applied to complicated test statistics where there is no existing formula for the standard error, such as the many test statistics employed in sequencing studies. It is worth noting that our procedure is computationally non-trivial due to the use of bootstrapping, but the runtime is reasonable for modern computers. For 1000 runs of the survival scenario with 100 bootstrap replications, for instance, the output of proc.time was 773.288 or about 13 seconds of elapsed time.

Figure 4.1: QQ Plot for the Unweighted Method

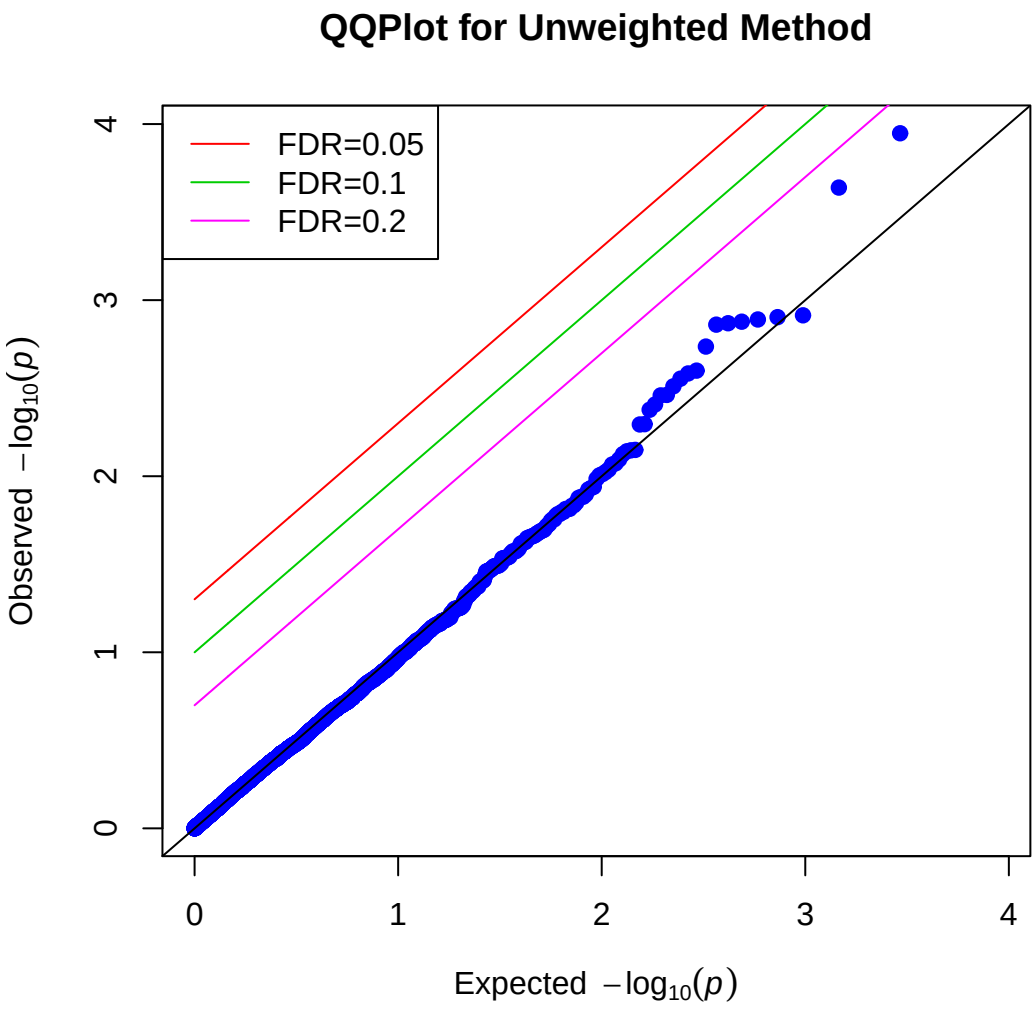

Figure 4.2: QQ Plot for the Weighted Bootstrap Method

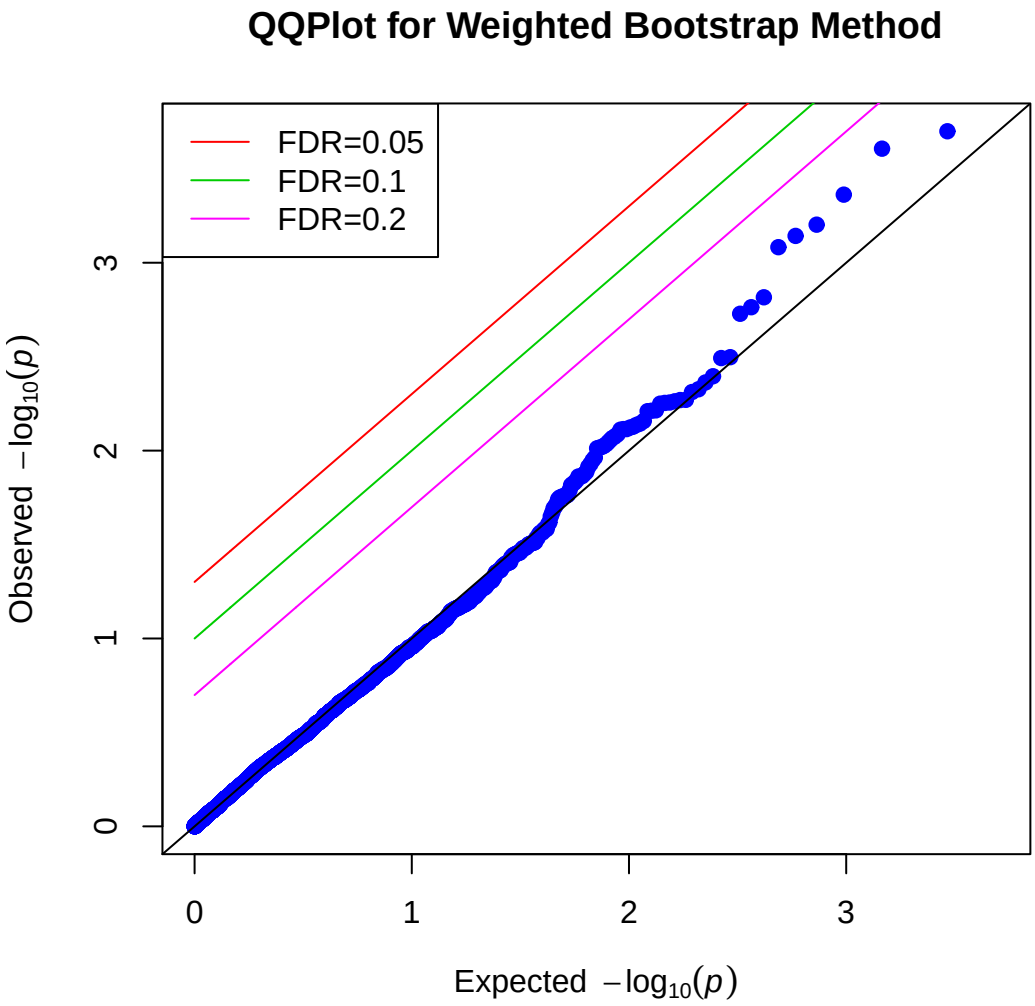

Table 4.2: Results of Simulations for Intermediate Ordinal Phenotypes

Method	Result			
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Naive (Bias)	-0.054	-0.068	-0.053	-0.069
Controls only (Bias)	0.075	0.051	0.063	0.093
Cases only (Bias)	0.805	-0.815	0.226	0.903
Adjusted for case status (Bias)	0.057	0.027	0.054	0.101
Bootstrap (Bias)	0.02	0.006	0.015	0.015
Naive (Coverage)	0.909	0.836	0.91	0.871
Controls only (Coverage)	0.937	0.937	0.948	0.933
Cases only (Coverage)	0	0	0	0
Adjusted for case status (Coverage)	0.904	0.937	0.913	0.824
Bootstrap (Coverage)	0.943	0.944	0.948	0.951
Bootstrap (Width)	0.519	0.399	0.477	0.512

Table 4.3: Results of Simulations for Intermediate Time-to-Event Phenotypes

Method	Result
Naive (Bias)	-0.457
Controls only (Bias)	0.272
Cases only (Bias)	-0.800
Adjusted for case status (Bias)	0.091
Bootstrap (Bias)	-0.017
Naive (Coverage)	0.006
Controls only (Coverage)	0.396
Cases only (Coverage)	0.225
Adjusted for case status (Coverage)	1.000
Bootstrap (Coverage)	0.944
Bootstrap (Width)	0.439

CHAPTER 5: SPARSE FACTOR ANALYSIS

5.1 Introduction

In many studies, we collect a large amout of data for each participant. The covariates may be largely correlated and repetitive. Thus, it may be of interest to reduce the covariates to a more manageable subset. Principal components analysis (PCA) and exploratory factor analysis (EFA) are two classical methods designed for this purpose. PCA calculates linear combinations of the variables which explain the maximal amout of variance in the factors. EFA, by contrast, assumes that the observed variables are a linear combination of a number of latent factors. However, the results of both PCA and EFA are usually very difficult to interpret and may not be as sparse as desired. It is desirable to have sparse loadings, meaning that most loadings are equal to zero and only a small subset are nonzero. Sparse loadings facilitate solutions that are more intuitive and make practical sense in real data applications.

Consequently, new methods have been proposed recently in the interest of sparsity. (157) proposed an iterative method for sparse principal component analysis based on elastic net. (147) proposed another iterative method for sparse principal component analysis that maximizes the percentage of variance explained using penalized matrix decomposition.

However, the aforementioned methods are not immediately applicable to the framework of factor analysis. Factor analysis is commonly used for psychometric data, when it is reasonable to assume that there are a small number of latent factors which drive the

observed variables and simple interpretations are of the utmost importance. Compared to principal component analysis, factor analysis by design sacrifices variance explained for interpretability. Similarly, the increase in variance explained by iterating the method of (147) may come with a decrease in interpretability.

Consider the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study, a large multicenter study of temporomandibular disorders (TMD). Researchers measured a large number of covariates related to orofacial pain. Within the quantitative sensory testing domain, (46) uses EFA to identify five underlying clinically meaningful components of the data. Similar analyses (76, 37) were performed for data on the autonomic nervous system function and psychosocial responses. Unfortunately, the loadings are not sparse for any of these domains. Investigators would like to be able to succintly explain latent factors that explain first-onset and chronic TMD, identify variables that do not add additional information after accounting for the other variables, and reduce the number of variables collected for follow-up studies.

We seek a method that generalizes EFA to have sparse loadings. However, because we would like to use the loadings to predict TMD-related outcomes, we want to make sure that the most important covariate information is not lost. In short, we propose a method of sparse factor analysis that is predictive of first-onset and chronic TMD, while minimizing the loss in variance explained compared to other methods. In section 5.2, we discuss our proposed method. In section 5.3, we show how the method performs in simulations. In section 5.4, we apply the method to the OPPERA study.

5.2 Proposed Methodology

5.2.1 Notation

This paper generalizes the method of (147) to the EFA setting with sparse loadings. Assume there are n subjects and p covariates. Let $X = (X_i, \ldots, X_n)'$ be an $n \times p$ matrix

containing the covariate information, $rank(X) = K \leq \min(n, p)$, and $E[X_i] = \mu_i$. We assume that there are k latent factors for each individual i, represented by the $k \times 1$ vector $F_i = (F_{1i}, \ldots, F_{ki})'$, F_{ji} is a 1-dimensional unobserved random variable for all $i = 1, \ldots, n$ and $j = 1, \ldots, k$, and $F = (F_1, \ldots, F_n)'$ is a $n \times k$ matrix of these factors. Often Also, $X - \mu = FL' + \epsilon$, where L is a $p \times k$ matrix of factor loadings, ϵ is a $n \times p$ matrix of random errors, $\mu = E[X]$, $E(\epsilon_{ij}) = 0$, $Var(\epsilon_{ij}) = \sigma_{ij}^2$ and $Cov(\epsilon_{ij}, \epsilon_{i',j'}) = 0$ as long as $i = i'$ and $j = j'$ are not both true. We assume that $E[F] = 0$, $Cov(F_i) = I$ for all i, and F and ϵ are independent.

In the OPPERA study, X is the matrix of all p variables collected on n individuals for a given domain, such as quantitative sensory testing (QST). The factors F denote a matrix of unobserved variables that affect our measurements in that domain. We seek to reduce the number of variables collected to a subset of k interpretable components that explain a maximal amount of variance in our measurements within that domain.

The singular value decomposition (SVD) of X is given by $X = UDV'$ where U is an $n \times n$ matrix, V is a $p \times p$ matrix, D is a $n \times p$ rectangular diagonal matrix $U'U = I_n$ and $V'V = I_p$. Denote u_j and v_j as the j^{th} columns of U and V, respectively. The notation $||M||_F$ denotes the Frobenius norm of a matrix M.

The soft threshholding function is given by $S(x, c) = sign(x)(|x| - c)_+$ where $y_+ = yI(y > 0)$. In other words, if possible, softthresholding reduces each coefficient by c in absolute value. Otherwise, it sets the coefficient to zero. Practically, this means that coefficients whose values are nonzero but too small to be practically meaningful are set exactly equal to zero.

5.2.2 Background Information

The optimization problem in (147) is to

$$\text{Minimize}_{d,u,v} ||X - duv'||_F^2 \text{ subject to } P_1(u) < c_1, P_2(v) < c_2, ||u||_2^2 = 1, ||v||_2^2 = 1, d > 0 \tag{5.1}$$

where P_1 and P_2 are convex penalty functions, u is a $n \times 1$ vector, and v is a $p \times 1$ vector. Typically the Lasso penalty, $P_1(u) = ||u||_1$ is used. We may rewrite the optimization using theorem 2.1 in (147), reproduced here:

Theorem 1. *Let U and V be $n \times K$ and $p \times K$ orthogonal matrices and D a diagonal matrix with diagonal elements $d_1, \ldots, d_k$. Then,*

$$\frac{1}{2}||X - UDV'||_F^2 = \frac{1}{2}||X||_F^2 - \sum_{k=1}^{K} u_k X v_k d_k + \frac{1}{2}\sum_{k=1}^{K} {d_k}^2$$

Applying Theorem 1 with $K = 1$ reduces problem (5.1) to

$$\text{Maximize}_u u'Xv \text{ subject to } P_1(u) < c_2, P_2(v) < c_2, ||u||_2^2 = 1, ||v||_2^2 = 1. \tag{5.2}$$

where the sole diagonal element $d_1 = d$ is positive. In order to have a convex solution, problem (5.2) is redefined as

$$\text{Maximize}_u u'Xv \text{ subject to } P_1(u) < c_2, P_2(v) < c_2, ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \tag{5.3}$$

Finally, (147) impose constant L-1 norm penalties on u and v to yield the following.

$$\text{Maximize } u'Xv \text{ subject to } ||v||_1 < c_2, ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \tag{5.4}$$

and

$$\text{Maximize } u'Xv \text{ subject to } ||u||_1 < c_1, ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \tag{5.5}$$

In short, for the first factor, they propose an iterative algorithm, solving (5.4) for u based on fixed v, solving (5.5) for v based on the esimate for u and repeating until convergence. Specifically, in (147), see Algorithms 1 and 2 for the single and multiple factor solutions, the special form given in Algorithm 3, and the notes about sparsity in section 3.2 for more details.

5.2.3 Our Algorithm

First, we center and scale X, resulting in the matrix X_s. We perform traditional factor analysis on X_s and extract the factor loadings $v_{f1}, \ldots, v_{fk}$ for k factors. We perform soft threshholding on the k factors and normalize the results to yield $v_1, \ldots, v_k$.

Next, for each factor, we update u based on X_s and v. Namely, $u_1 = X_s v_1 / ||X_s v_1||$, and for $j = 2, \ldots, k$, $u_j = \epsilon_j$ where $\tilde{y}_j = X_s v_j$, $\tilde{X}_j = (u_1, \ldots, u_{j-1})$ denotes the first $j-1$ columns of U, and u_j is the set of residuals from the least squares equation

$$\tilde{y}_j = \tilde{X}_j \beta_j + \epsilon_j \tag{5.6}$$

The following algorithm summarizes the method.

1. Find the initial estimate for v.
   (a) Scale and center X to yield X_s.
   (b) Perform traditional factor analysis on X_s.
   (c) Extract the factor loadings $v_{f1}, \ldots, v_{fk}$ for k factors.
   (d) $v_j = \frac{S(v_{fj}, \Delta_j)}{||S(v_{fj}, \Delta_j)||_2}$ for $j = 1, \ldots, k$
2. Update u based on v.

(a) $u_1 \leftarrow \frac{X_s v_1}{||X_s v_1||_2}$

(b) for $j = 2, \ldots, k$, $u_j = \epsilon_j$ is the set of residuals from the fit of the data to the model in (5.6).

Note that the thresholds, $\Delta_1, \ldots, \Delta_k$, may vary for each component.

Finally, we use the formula in (109) to calculate the percentage of variance explained by this procedure. Namely, the cumulative percentage of variance explained by first j components is given by $tr({X_j}'X_j)$ where $V_j = [v_1, \ldots, v_j]$ is the $p \times j$ matrix of the first j loading vectors and $X_j = X_s V_j (V_j' V_j)^{-1} V_j'$ is the $n \times j$ projection of X_s onto the subspace generated by V_j.

Unlike the method of (147), our procedure does not iterate between these steps. Consequently, the percentage of variance explained will be less than the variance explained by their method. However, the trade off is that the results of our procedure should be more sparse and lend to more concise interpretations, which are of interest in the factor analysis setting. In addition, lack of iteration results in greater computational efficiency.

5.3 Simulations

This section describes simulations that demonstrate our method and compare it to a number of alternatives. Samples of size $n = 1000$ were generated with $p = 30$ covariates. For each individual, $k = 5$ factors were generated as independent normal random variables with mean zero and standard deviations 10, 20, 30, 30, 30. That is, for all i, $F_i = (F_{1i}, \ldots, F_{5i})'$ is a 5×1 random normal variable with $E[F_{ji}] = 0$ and $Var[F_i] = Diag(100, 400, 900, 900, 900)$. This corresponds to having one factor with a small amount of noise, a second factor with increased noise, and 3 additional factors that are noisier than the first two. The full matrix of all factors is given by the $n \times k$ matrix $F = (F_1, \ldots, F_n)'$.

We generated the $p \times k$ normally distributed loading matrix L. Loadings were independent and each had standard deviation 0.1. The mean $\mu_{ij} = E[L_{ij}]$ varied based on the row i and column j of the matrix L. Namely, $\mu_{ij} = 0.8$ if $j = 1$ and $i = 1, ...6$; $\mu_{ij} = 0.7$ if $j = 2$ and $i = 7, \ldots, 12$; $\mu_{ij} = 0.6$ if $j = 3$, and $i = 13, ...18$; $\mu_{ij} = 0.5$ if $j = 4$, $i = 19, ..., 24$; or $\mu_{ij} = 0.4$ if $j = 5$, and $i = 25, ..., 30$. In other words, the first 6 elements of column 1 have mean 0.8, the second 6 elements of column 2 have mean 0.7, the third six elements of column 3 have mean 0.6, the fourth six elements of column 4 have mean 0.5, and the last six elements of column 5 have mean 0.4; all other elements have mean zero. The standard deviation of 0.1 for all elements means that all of the loadings are nonzero, some are essentially just noise, and others are larger and more meaningful but have a small amount of noise as well.

The elements of the $n \times 1$ error vector ϵ were also independent and normally distributed with mean 0, variance 1. As in the methodology section, $X = FL' + \epsilon$.

For each individual, a binary outcome Y_i was generated based on the scaled factors, F_s. We used

$$\text{logit}\, P(Y = 1) = \tilde{F}\alpha$$

where α is a $(k + 1) \times 1$ vector of ones, F_s is an $n \times k$ matrix of the factors, scaled to have mean zero and unit variance, and $\tilde{F} = (1, F_s)$ is a $n \times (k + 1)$ matrix with the first column consisting of all 1's and the remaining entries identical to F_s. Here, $Y = (Y_1, \ldots, Y_n)'$.

We compared the performance of our method with traditional factor analysis, three applications of (147), and two applications of (157). The different applications of the competing methods varied based on level of sparsity specified. We used $c(0.06, 0.16, 0.1, 0.5, 0.5)$ and $c(7, 4, 4, 1, 1)$ in the (157) function which corresponded to low and high sparsity, respectively. The (147) applications varied based on 1, $\sqrt(p)/2$, $\sqrt(p)$, which correspond to high, medium, and low sparsity, where p denotes the number of

variables under consideration. For our method, we also investigated softthreshold values of 0.4, 0.5, 0.6, 0.7, and 0.8 for all components, as well as one scenario with increasing thresholds for each components and one scenario with decreasing thresholds for the components.

Simulations were run independently 1000 times. For each run, we fit logistic regression models of the outcome based on each component separately for each method and recorded the parameter estimate and corresponding p-value. We recorded the average percent variance explained, and average parameter estimate (Table 5.1) and average p-value (Table 5.2) for each of the five components. We also noted the number of variables with nonzero loadings for each component and each method (Table 5.3). We refer to our proposed method as Sparse Factor Analysis (SFA).

Our method captured a large percentage of the variance, around 71-82%, depending on the thresholds used. Traditional factor analysis and the non-sparse methods of (147) and (157) explained 91%. Sparse applications of (147) and (157) explained far less variance (<40%).

For all thresholds, parameter estimates for all components were comparable in magnitude to those for factor analysis but much larger than for the competing methods. P-values were strongly significant for all methods. This indicates that the factor scores from our method were more strongly related to the outcome than the competing method scores were to the outcome. The associations were positive for our method, as intended, and often negative for the other methods. Moreover, our method was clearly more sparse than traditional factor analysis and all but the most sparse versions of (147) and (157). We retained an average of about 6 variables per component with nonzero loadings, compared to all or nearly all of the 30 variables for these other methods. This was favorable, as we had generated only 6 variables to be meaningful for each component and generated all others to represent noise.

In conclusion, the simulations show that for various thresholds, our method results in sparse factor scores that still explain a majority of the variance and are correlated with the outcome of interest. These were exactly the properties that were desired from a practical standpoint.

5.4 Data Application

This section details the application to the OPPERA study. The OPPERA study is comprised of a baseline case-control study (113) of chronic temporomandibular disorders (TMD) and a prospective cohort study (1) of first-onset TMD. The prospective cohort study consists of 3258 subjects who were free from TMD at basline. Officially, the baseline case-control data consists of half of the controls, i.e. 1633 subjects who were free from TMD at baseline, as well as 185 previously diagnosed cases of TMD. However, for purposes of measuring predictability of chronic TMD, it makes sense to consider not only the 1633 randomly selected controls, but the entire collection of 3258 controls as well as the 185 cases. This yields the total sample size in OPPERA, $n = 3443$. Within the prospective cohort, 2737 subjects had some follow-up data.

The goals of this analysis are to see which variables are most important in predicting chronic and first-onset TMD, to establish a subset of the variables that may adequately explain TMD status, and to see if the use of different thresholds in our methodology changes the results. Percentage of variance explained is far less important than the combination of interpretability and the predictive value of the model. Predictive value of the components was examined via logistic regression models. For each component, a model was fit to the cohort against the outcome. The outcomes of interest, chronic TMD and first-onset TMD, are defined in (113) and (1). The parameter estimate, standard error, odds ratio, and p-values were recorded and compared between various applications of the methods in this paper and the results of traditional factor analysis.

This paper presents results for the case control subjects and prospective cohort of the OPPERA study described above. Whenever possible, we consider three domains of the OPPERA study: quantitative sensory testing (QST), autonomic nervous system function, and psychosocial factors. Details on the these domains are provided in (75). Previous results (46, 76, 37) indicate that 5, 5, and 4 components, respectively are optimal for these datasets. Results in this paper use the same number of components as the authors did in the previous results. First, cases and controls are analyzed separately when possible. Next, factor scores are generated for the entire OPPERA cohort, consisting of both cases and controls. Finally, the results are compared for predictability of chronic TMD and first-onset TMD.

5.4.1 The Baseline Case-Control Study of Chronic TMD

Intoduction

This portion of the paper discusses the loadings for cases and controls separately for the QST dataset. For the autonomic dataset, the cases and controls were analyzed together. The psychosocial domain was not considered because there was not a dataset consisting of only the cases and controls in the official baseline case-control study as discussed in (37).

QST Results

This baseline case-control data includes 1633 controls and 185 cases with information on $p = 33$ variables measuring quantitative sensory testing. Loadings were generated for cases and controls separately to facilitate comparison with the results of (46). For these separate analyses, the only soft threshold considered for all components was 0.4, as is commonly used as an arbitrary cut-off in factor analysis.

Results in this first analysis were compared with competing methods. Our method

explained 66.4% of the variance in the data for controls. By comparison, the (147) method explained between 15.1 % and 71.9% of the variance, depending on the sparsity specified. The (157) method also explained between 31.8% and 71.9% of the variance.

Our sparse factors were highly correlated with the original factor scores. Suppose that F_0 represents the factor scores based on EFA, and F represents the factor scores from our method. The diagonal of $Cor(F_0, F)$ is $d = (0.934, 0.945, 0.950, 0.977, 0.995)$.

Table 5.4 shows the loadings from our method. Compared to the results in (46), the loadings in the present paper are smaller but occur in the same places. This means sparse factor analysis uncovered the same components as standard factor analysis, but the sparse factor analysis method rigorously achieved truly sparse loadings rather than ignoring or artificially setting small loadings to zero. These components are known as "heat pain ratings" (component 1), "heat pain aftersensations and tolerance" (component 2), "mechanical cutaneous pain sensitivity" (component 3), "pressure pain thresholds" (component 4), and "heat pain temporal summation (TS)" (component 5). Results for cases, found in Table A2.14, are similar. The clinical interpretation of these components is discussed further in (46)

Autonomic Results

This dataset consists of 1630 cases and 185 controls with 42 measured variables on autonomic function.

In this analysis and for subsequent sections, we applied our method multiple times with various soft thresholds. Here, we examine thresholds of 0.4 to 0.8 for all components. Thresholds were identical for each component for results in this section. Using thresholds of 0.4 for all 5 components yielded similar components to that in Maixner et al. (76), but with true zero loadings which previously were only arbitrarily set to zero. The components were Blood Pressure, Stroop Heart Rate Variability (HRV), Heart Rate,

Resting HRV and Orthostatic HRV. See Maixner et al. (76) for a biological description of these components.

As expected, increasing the threshold from 0.4 to 0.8 for all components increased the sparsity. When we used a threshold of 0.8 for all components, there were only 24 variables with a nonzero loading on at least one component (Table A2.18). For example, average systolic blood pressure and heart rate were no longer needed in the model after measuring average resting systolic blood pressure and mean arterial blood pressure (MAP). Similarly, the initial orthostatic measurements (SBP, DBP, MAP, and HR) were no longer needed. By contrast, traditional factor analysis and the application of our method with the threshold of 0.4 for all components yielded all 42 variables with nonzero loadings.

Despite the increased sparsity, overall predictability remained similar for the thresholds as shown in Tables 5.5-5.7, which report the parameter estimates and their p-values for three components at various thresholds. The second component was highly predictive of chronic TMD for thresholds up to 0.7 and then significant but less strongly so for a threshold of 0.8. However, even at this high threshold, our component was more strongly predictive of chronic TMD than was the corresponding component from traditional factor analysis. Moreover, the third component was highly significant for all thresholds. The fourth component was not predictive of chronic TMD using the methods in this paper for any of the thresholds explored. While the fourth component of a traditional factor analysis was predictive, it was only marginally so ($p=0.0362$), and as the threshold increased, our parameters increased in magnitude and the p-values dropped. Curiously, higher thresholds resulted in higher parameter estimates (in magnitude) for components two and three, which were associated with chronic TMD. This indicates that the components were more strongly predictive of TMD once soft thresholding was implemented with higher thresholds. The first and fifth components

are not discussed because they were not predictive of chronic TMD for either method. In short, it can be argued that soft thresholding did not reduce the predictability of the model for chronic TMD, even when we dramatically reduced the number of variables under consideration within the autonomic domain.

We report thus the loadings for thresholds of 0.7 for all components in Table 5.8. For the other thresholds, see Tables A2.15-A2.18.

5.4.2 Application to the Full OPPERA Cohort

For the full dataset, the same scores were reported in the context of both chronic TMD and first-onset TMD. It is only the predictability that changes. The following subsections detail the factor scores in general as well as their relationship to chronic and first-onset TMD.

QST Loadings for the Full Cohort

The full OPPERA data consists of $n = 3443$ subjects with information on $p = 33$ variables measuring quantitative sensory testing.

Table 5.9 shows the factor loadings when our method is applied to this QST data with a threshold of 0.8 for all components. Additional results are given for various thresholds in Tables A2.1-A2.4. When the threshold of $c = 0.4$ was used for all components, our method explained 68.9% of the variance, compared to 73.4% for standard factor analysis. Higher thresholds, such as 0.8 for all components, correspond to decreased variance explained, down to 44%. Fortunately, as discussed in sections 5.4.2 and 5.4.2, this does not negatively impact predictability with respect to the outcomes of interest.

Loadings were qualitatively similar for each threshold examined with one logical but notable exception. Larger thresholds resulted in fewer variables with nonzero

loadings. For example, the mechanical cutaneous pain sensitivity component only had one variable with a nonzero loading: the rating 15 seconds after being probed by a 256 mN probe. Additionally, the heat pain temporal summation component only had 3 nonzero variables, one of which was very small (0.08). By contrast, all 8 variables in the mechanical cutaneous pain sensitivity component and all 6 variables in the heat pain temporal summation component loaded nonzero (and exceeded 0.4) in (46).

Application to the QST Data for Chronic TMD

For standard factor analysis, the third and fourth component were predictive of chronic TMD. The first four factors (i.e. all but heat pain temporal summation) for our method were all highly predictive of chronic TMD. Moreover, especially for higher thresholds, our loadings were sparse. Consider Table 5.9, which shows loadings when we use a threshold of 0.8 for all components. The fourth component, measuring mechanical cutaneous pain sensitivity, was weighted less heavily than previously. In fact, the only variable remaining in this component was the after sensation rating (15 s, 256 mN probe). The fifth component, temporal summation, was also not weighted as heavily, nor was it predictive of chronic TMD.

Application to the QST Data for First-Onset TMD

None of the components were significantly associated with TMD, but none of the components of a traditional EFA/PCA analysis were significantly associated with TMD either. Thus, our method was no worse than the traditional method at predicting first-onset TMD.

Results remained qualitatively similar for different threshold values. Decreasing the threshold values rendered heat pain tolerance and the temporal summation single stimulus variables nonzero in all components. For thresholds of 0.7 or 0.8 for all

components, the temporal summation slope of the line for 46 degrees was set equal to zero. When the threshold was 0.8 for all components, single stimulus and temporal summation for either probe, ratings 30 seconds after exposure to either probe or 15 seconds after the 512 mN probe, Single Stimulus 46 and 50 degrees and all three temporal summation slope variables were also zero. This left the after Sensation Rating (15 s, 256 mN probe) as the sole nonzero loading on the 4th component and the two remaining temporal summation slope variables as the only nonzero loadings on the 5th component when $c = 0.8$. All other variables only had minor changes in the loadings as the thresholds changed.

Autonomic Loadings for the Follow-up Cohort

Loadings are provided in Table 5.10 for the 2737 individuals in the prospective cohort study with at least 1 QHU. This allowed examination of associations with first-onset TMD. Because predictability in the case-control study was already examined for those in the official base-line case control study, and because of the plethora of tables already in this paper, additional analysis was not done to examine chronic TMD in the larger cohort. Moreover, these results should be virtually identical as they were in the multiple subsets analyzed for the QST domain. If desired, this additional analysis can be added before we publish this paper.

Application to the Autonomic Data for First-Onset TMD

There are 42 variables measured in the autonomic domain. (76) previously fit traditional EFA/PCA to the autonomic baseline case-control data. This paper applies the SFA method to the autonomic follow-up cohort data. SFA explained 71.1% of the data, compared to 74.7% for the EFA/PCA analysis. Moreover, the first component (blood pressure) scores were predictive of first-onset TMD. We fit the model using

$c = 0.4, 0.5, 0.6, 0.7, 0.8$. Results were quantitatively similar, with only baseline heart rate variability and heart rate variability related to the Stroop questionnaire having all components set to zero. The percent variance explained dropped very slightly when the threshold for all components increased.

Predictive accuracy remained high for all thresholds but decreased as the threshold decreased. In particular, logistic regression models were fit for each set of factor scores to predict TMD. Parameter estimates and odds ratios were compared for the different threshold values and for traditional factor analysis. As shown in Table 5.11, the parameter estimates were higher for the sparse factor analysis method than for the traditional factor analysis when $c = 0.4$ or $c = 0.5$. For higher thresholds, the relationship reversed. Yet, differences were slight for all thresholds. Odds ratios were similar for method compared to the traditional method and only dropped noticeably for the highest threshold. P-values only increased appreciably for the two highest threshold values, but were still under 0.05. This indicates that even with sparse loadings, the blood pressure component is predictive of TMD.

Loadings for the Psychosocial Data

For the psychosocial domain, results were applied to the full dataset of 3443 participants. The psychosocial data consists of 21 variables as previously analyzed in (37). Loadings for the psychosocial dataset are provided in Table 5.12 for when the threshold is 0.6 for all components. Other loadings may be found in the appendix in Tables A2.19, A2.20, and A2.21. In fact, for the psychosocial domain, we considered having different thresholds for different components. See Tables A2.22-A2.24.

The components were Global Psychological Symptoms, Stress and Negative Affectivity, Passive Pain Coping, and Active Pain Coping. These are the same components identified in (37), except that the order of the first two components were swtiched. For a more

detailed description of their meaning, please see (37).

Application to the Psychosocial Data for Chronic TMD

The first, third, and fourth psychosocial components were predictive of chronic TMD regardless of the threshold. That is, global psychological symptoms, and both active and passive pain coping were associated with chronic TMD. The magnitude of the parameter estimate and p-value did not have a monotonic relationship with the thresholds. In fact, when the threshold increased, the parameter estimate increased and the p-value decreased for component 3, passive pain coping. This indicates that the larger thresholds both made the loadings more sparse and increased the predictability compared to traditional factor analysis.

Application to the Psychosocial Data for First-Onset TMD

The first two components, Global Psychological Symptoms, Stress and Negative Affectivity, were predictive of first-onset TMD for all thresholds of our method as well as for standard EFA. However, unlike ERA, the loadings using the method proposed in this paper are sparse especially for larger thresholds. Notable variables that had zero loadings for all components included KOHN global score, POMS Negative Affective Score, CSQ Praying, EPQ N scale and EPQ E scale. In short, our method produced sparse components that were predictive of first-onset TMD.

5.5 Discussion

This work proposes a method for factor analysis that allows for sparse factors, motivated by the research questions in the OPPERA study. The use of soft thresholding is employed to shrink coefficients, followed by a simple one step update. Our method is similar to the sparse principal component analysis method of (147) but does not iterate.

This results in increased computational efficiency and sparsity, while maintaining a reasonable percentage of variance explained. Simulations showed that promising results, namely, that even as the thresholds increase for one or more components, the parameters are still significantly associated with the outcomes of interest.

We applied the method to various domains within the OPPERA study. Compared to previous results, our loadings were more sparse yet remained predictive of both chronic and first-onset TMD, where applicable. Moreover, the sparse results of our method illuminate to investigators variables which may be redundant and may not be needed in follow-up studies.

The method proposed in this paper is preferred to other methods, such as (147) and (157), which are designed for principal component analysis. While factor analysis may not be as desirable statistically, it provides much more easily interpretable components. Thus, it is commonly used in the psychometric literature. A generalization such as ours would improve their research by allowing them to consider sparse factors.

Table 5.1: Parameters for Different Thresholds and Other Methods

Method Used	PVE	Parameter for Components				
		1	2	3	4	5
SFA (0.8,...,0.8)	0.7085	0.6169	0.6186	0.6194	0.6053	0.5911
SFA (0.7,...,0.7)	0.7750	0.6179	0.6220	0.6246	0.6017	0.5901
SFA (0.6,...,0.6)	0.7987	0.6191	0.6253	0.6293	0.5965	0.5870
SFA (0.5,...,0.5)	0.8110	0.6203	0.6287	0.6331	0.5914	0.5838
SFA (0.4,...,0.4)	0.8207	0.6211	0.6317	0.6360	0.5871	0.5809
SFA (0.4,...,0.8)	0.7721	0.6211	0.6289	0.6284	0.5921	0.5829
SFA (0.8,...,0.4)	0.7955	0.6169	0.6219	0.6298	0.6006	0.5890
EFA	0.9054	0.6169	0.6173	0.6224	0.6076	0.6044
Witten (Sparse)	0.1667	-0.0828	0.0126	0.1157	0.1568	0.0145
Witten (Full)	0.9054	-0.1970	0.0068	0.1787	0.2002	0.0697
Witten (Medium)	0.8564	-0.0707	0.0021	0.1219	0.1610	-0.0139
Zou (Full)	0.9054	-0.0300	-0.0061	-0.0224	-0.0017	-0.0178
Zou (Sparse)	0.3752	-0.0341	0.0001	0.0094	0.0168	-0.0126

Table 5.2: P-Values for Different Thresholds and Other Methods

Method Used	P-value for Components				
	1	2	3	4	5
SFA (0.8,...,0.8)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
SFA (0.7,...,0.7)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
SFA (0.6,...,0.6)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
SFA (0.5,...,0.5)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
SFA (0.4,...,0.4)	<0.0001	0.0001	<0.0001	<0.0001	<0.0001
SFA (0.4,...,0.8)	<0.0001	0.0001	<0.0001	<0.0001	<0.0001
SFA (0.8,...,0.4)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
EFA	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Witten (Sparse)	0.0192	0.0096	0.0138	0.0112	0.0116
Witten (Full)	0.0572	0.0557	0.0590	0.0630	0.0814
Witten (Medium)	0.0005	<0.0001	<0.0001	<0.0001	<0.0001
Zou (Full)	0.0572	0.0557	0.0590	0.0630	0.0814
Zou (Sparse)	0.0103	0.0041	0.0023	0.0065	0.0086

Table 5.3: Number of Variables Retained for Different Thresholds and Other Methods

Method Used	Variables Retained for Components					
	1	2	3	4	5	Total
SFA (0.8,...,0.8)	5.919	5.761	5.468	5.203	5.080	27.431
SFA (0.7,...,0.7)	6.030	6.009	5.934	5.761	5.699	29.433
SFA (0.6,...,0.6)	6.115	6.120	6.111	5.992	5.911	29.914
SFA (0.5,...,0.5)	6.318	6.256	6.268	6.179	6.051	29.992
SFA (0.4,...,0.4)	6.896	6.637	6.585	6.432	6.207	29.998
SFA (0.4,...,0.8)	6.896	6.256	6.111	5.761	5.080	29.420
SFA (0.8,...,0.4)	5.919	6.009	6.111	6.179	6.207	29.893
EFA	30.000	30.000	30.000	30.000	30.000	30.000
Witten (Sparse)	1.000	1.000	1.000	1.000	1.000	5.000
Witten (Full)	30.000	30.000	30.000	30.000	30.000	30.000
Witten (Medium)	10.900	11.820	12.612	13.638	14.193	30.000
Zou (Full)	29.973	29.925	29.956	29.747	29.741	30.000
Zou (Sparse)	7.000	4.000	4.000	1.000	1.000	16.986

Table 5.4: PCA/SFA QST Results for Controls, c=0.4

Component	1	2	3	4	5
Pressure Pain Threshhold					
Temporalis	0	0	0	0.47	0
Masseter	0	0	0	0.50	0
TMJ	0	0	0	0.47	0
Trapezius	0	0	0	0.39	0
Epicondyl	0	0	0	0.39	0
Mechanical Cutaneous Pain Threshhold	0	0	-0.01	0	0
Single Stimulus					
256 mN Probe	0	0	0.34	0	0
512 mN Probe	0	0	0.35	0	0
After Sensation Ratings					
15 s, 256 mN probe	0	0	0.48	0	0
30 s, 256 mN probe	0	0	0.41	0	0
15 s, 512 mN probe	0	0.00	0.42	0	0
30 s, 512 mN probe	0	0.01	0.38	0	0
Temporal Summation					
256 mN probe	0	0	0.19	0	0
512 mN probe	0	0	0	0	0
Heat Pain Tolerance	-0.05	0	0	0	0
Single Stimulus					
46	0.33	0	0	0	0
48	0.42	0	0	0	0
50	0.43	0	0	0	0
Area Under the Curve					
46	0.42	0	0	0	0
48	0.45	0	0	0	0
50	0.38	0	0	0	0
After Sensations					
15 s, 46	0	0.36	0	0	0
30 s, 46	0	0.40	0	0	0
15 s, 48	0	0.41	0	0	0
30 s, 48	0	0.44	0	0	0
15 s, 50	0	0.41	0	0	0
30 s, 50	0	0.42	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.47
48	0	0	0	0	0.50
50	0	0	0	0	0.32
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0.33
48	0	0	0	0	0.44
50	0	0	0	0	0.36
Percent Variance Explained	0.15	0.30	0.44	0.55	0.66

Table 5.5: Autonomic Parameter Estimates and Different Thresholds, Component 2

Method Details	Parameter Estimate	Standard Error	Odds Ratio	P-value
Traditional EFA/PCA	-0.20445	0.09420	0.8151	0.03
SFA $c = 0.8$	-0.25411	0.10456	0.7756	0.0151
SFA $c = 0.7$	-0.30642	0.09836	0.7361	0.00184
SFA $c = 0.6$	-0.31278	0.09455	0.7314	0.000939
SFA $c = 0.5$	-0.31401	0.09242	0.7305	0.00068
SFA $c = 0.4$	-0.31515	0.09112	0.7297	0.000543

Table 5.6: Autonomic Parameter Estimates and Different Thresholds, Component 3

Method Details	Parameter Estimate	Standard Error	Odds Ratio	P-value
Traditional EFA/PCA	0.23284	0.07621	1.2622	0.00225
SFA $c = 0.8$	0.28326	0.07356	1.3275	0.000118
SFA $c = 0.7$	0.24390	0.07376	1.2762	0.000944
SFA $c = 0.6$	0.22576	0.07402	1.2533	0.00229
SFA $c = 0.5$	0.21577	0.07422	1.2408	0.00365
SFA $c = 0.4$	0.20941	0.07436	1.2330	0.00486

Table 5.7: Autonomic Parameter Estimates and Different Thresholds, Component 4

Method Details	Parameter Estimate	Standard Error	Odds Ratio	P-value
Traditional EFA/PCA	-0.16051	0.07664	0.8517	0.0362
SFA $c = 0.8$	-0.10838	0.07626	0.8973	0.155
SFA $c = 0.7$	-0.09100	0.07658	0.9130	0.235
SFA $c = 0.6$	-0.08228	0.07673	0.9210	0.284
SFA $c = 0.5$	-0.07899	0.07695	0.9240	0.305
SFA $c = 0.4$	-0.07664	0.07714	0.9262	0.32

Table 5.8: Autonomic Results for the OPPERA Case Control Study, c=0.7

Component	1	2	3	4	5
AvgRestingSPB	0.25	0	0	0	0
AvgRestingDPB	0.07	0	0	0	0
AvgRestingMAP	0.27	0	0	0	0
AvgRestingHR	0	0	0.15	0	0
InitialOrthoSPB	0	0	0	0	0
InitialOrthoDPB	0.03	0	0	0	0
InitialOrthoMAP	0.17	0	0	0	0
InitialOrthoHR	0	0	0	0	0
StroopColorSPBmean	0.38	0	0	0	0
StroopColorDPBmean	0.31	0	0	0	0
StroopColorMAPmean	0.44	0	0	0	0
StroopColorHRmean	0	0	0.44	0	0
StroopEmotionalSPBmean	0.37	0	0	0	0
StroopEmotionalDPBmean	0.27	0	0	0	0
StroopEmotionalMAPmean	0.43	0	0	0	0
StroopEmotionalHRmean	0	0	0.48	0	0
QST.SHRVBaseline20MinMeanHR	0	0	0.14	0	0
QST.SHRVBaseline20MinSDNN	0	0	0	0.24	0
QST.SHRVBaseline20MinRMSSD	0	0	0	0	0
HRVBaseline20minLnTP	0	0	0	0.67	0
HRVBaseline20minLnVLF	0	0	0	0.24	0
HRVBaseline20minLnLF	0	0	0	0.55	0
HRVBaseline20minLnHF	0	0	0	0.36	0
QST.SHRVOrthostatic5MinMeanHR	0	0	0.26	0	0
QST.SHRVOrthostatic5MinSDNN	0	0	0	0	0.59
QST.SHRVOrthostatic5MinRMSSD	0	0	0	0	0.39
HRVOrtho5minLnTP	0	0	0	0	0.46
HRVOrtho5minLnVLF	0	0	0	0	0.04
HRVOrtho5minLnLF	0	0	0	0	0.25
HRVOrtho5minLnHF	0	0	0	0	0.47
QST.SHRVSTROOPColor5MinMeanHR	0	0	0.47	0	0
QST.SHRVSTROOPColor5MinSDNN	0	0.58	0	0	0
QST.SHRVSTROOPColor5MinRMSSD	0	0.53	0	0	0
HRVStroopColor5minLnTP	0	0.31	0	0	0
HRVStroopColor5minLnVLF	0	0	0	0	0
HRVStroopColor5minLnLF	0	0.17	0	0	0
0 HRVStroopColor5minLnHF	0	0.27	0	0	0
QST.SHRVSTROOPEmotion5mnMeanHR	0	0	0.51	0	0
QST.SHRVSTROOPEmotion5mnSDNN	0	0.35	0	0	0
QST.SHRVSTROOPEmotion5mnRMSSD	0	0.25	0	0	0
HRVStroopEmotion5minLnTP	0	0	0	0	0
HRVStroopEmotion5minLnVLF	0	0	0	0	0
Percent Variance Explained	0.17	0.29	0.41	0.5	0.59

Table 5.9: PCA/SFA QST Results for the Entire OPPERA Cohort, c=0.8

Component	1	2	3	4	5
Pressure Pain Threshold					
Temporalis	0	0	0.45	0	0
Masseter	0	0	0.6	0	0
TMJ	0	0	0.48	0	0
Trapezius	0	0	0.39	0	0
Epicondyl	0	0	0.11	0	0
Mechanical Cutaneous Pain Threshold	0	0	0.2	0	0
Single Stimulus					
256 mN Probe	0	0	0	0	0
512 mN Probe	0	0	0	0	0
After Sensation Ratings					
15 s, 256 mN probe	0	0	0	1	0
15 s, 512 mN probe	0	0	0	0	0
30 s, 256 mN probe	0	0	0	0	0
30 s, 512 mN probe	0	0	0	0	0
Temporal Summation					
256 mN probe	0	0	0	0	0
512 mN probe	0	0	0	0	0
Heat Pain Tolerance	0	0	0	0	0
Single Stimulus					
46	0	0	0	0	0
48	0.4	0	0	0	0
50	0.48	0	0	0	0
Area Under the Curve					
46	0.43	0	0	0	0
48	0.63	0	0	0	0
50	0.17	0	0	0	0
After Sensations					
15 s, 46	0	0.08	0	0	0
15 s, 48	0	0.39	0	0	0
15 s, 50	0	0.3	0	0	0
30 s, 46	0	0.31	0	0	0
30 s, 48	0	0.63	0	0	0
30 s, 50	0	0.52	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.39
48	0	0	0	0	0.92
50	0	0	0	0	0
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0
48	0	0	0	0	0.08
50	0	0	0	0	0
Percent Variance Explained	0.12	0.24	0.36	0.39	0.44

Table 5.10: PCA/SFA Autonomic Results for the OPPERA Followup Cohort, c=0.7

Component	1	2	3	4	5
AvgRestingSPB	0.27	0	0	0	0
AvgRestingDPB	0.06	0	0	0	0
AvgRestingMAP	0.27	0	0	0	0
AvgRestingHR	0	0	0.16	0	0
InitialOrthoSPB	0	0	0	0	0
InitialOrthoDPB	0.02	0	0	0	0
InitialOrthoMAP	0.18	0	0	0	0
InitialOrthoHR	0	0	0	0	0
StroopColorSPBmean	0.37	0	0	0	0
StroopColorDPBmean	0.31	0	0	0	0
StroopColorMAPmean	0.44	0	0	0	0
StroopColorHRmean	0	0	0.44	0	0
StroopEmotionalSPBmean	0.36	0	0	0	0
StroopEmotionalDPBmean	0.27	0	0	0	0
StroopEmotionalMAPmean	0.43	0	0	0	0
StroopEmotionalHRmean	0	0	0.48	0	0
QST.SHRVBaseline20MinMeanHR	0	0	0.14	0	0
QST.SHRVBaseline20MinSDNN	0	0	0	0.05	0
QST.SHRVBaseline20MinRMSSD	0	0	0	0	0
HRVBaseline20minLnTP	0	0	0	0.7	0
HRVBaseline20minLnVLF	0	0	0	0.27	0
HRVBaseline20minLnLF	0	0	0	0.56	0
HRVBaseline20minLnHF	0	0	0	0.34	0
QST.SHRVOrthostatic5MinMeanHR	0	0	0.25	0	0
QST.SHRVOrthostatic5MinSDNN	0	0	0	0	0.59
QST.SHRVOrthostatic5MinRMSSD	0	0	0	0	0.27
HRVOrtho5minLnTP	0	0	0	0	0.53
HRVOrtho5minLnVLF	0	0	0	0	0.04
HRVOrtho5minLnLF	0	0	0	0	0.25
HRVOrtho5minLnHF	0	0	0	0	0.47
QST.SHRVSTROOPColor5MinMeanHR	0	0	0.47	0	0
QST.SHRVSTROOPColor5MinSDNN	0	0.56	0	0	0
QST.SHRVSTROOPColor5MinRMSSD	0	0.44	0	0	0
HRVStroopColor5minLnTP	0	0.26	0	0	0
HRVStroopColor5minLnVLF	0	0	0	0	0
HRVStroopColor5minLnLF	0	0.11	0	0	0
HRVStroopColor5minLnHF	0	0.27	0	0	0
QST.SHRVSTROOPEmotion5mnMeanHR	0	0	0.51	0	0
QST.SHRVSTROOPEmotion5mnSDNN	0	0.44	0	0	0
QST.SHRVSTROOPEmotion5mnRMSSD	0	0.39	0	0	0
HRVStroopEmotion5minLnTP	0	0.06	0	0	0
HRVStroopEmotion5minLnVLF	0	0	0	0	0
Percent Variance Explained	0.17	0.29	0.41	0.49	0.58

Table 5.11: Autonomic Parameter Estimates and Different Thresholds for Component 1, First-Onset TMD

Method Details	Parameter Estimate	Standard Error	Odds Ratio	P-value
SFA $c = 0.8$	0.14086	0.0629	1.15	0.0251
SFA $c = 0.7$	0.1540	0.0627	1.17	0.0140
SFA $c = 0.6$	0.1642	0.0626	1.18	0.0087
SFA $c = 0.5$	0.1682	0.0625	1.18	0.0072
SFA $c = 0.4$	0.1704	0.0625	1.19	0.0064
Traditional EFA/PCA	0.1649	0.0626	1.18	0.0085

Table 5.12: PCA/SFA Psychosocial Results for the Entire OPPERA Cohort, c=0.6

Component	1	2	3	4
KOHN.Global.Score	0	0	0	0
PILL.Global.Score	0.09	0	0	0
POMS.PositiveAffectScore	0	-0.62	0	0
POMS.NegativeAffectScore	0	0	0	0
PSS.PerceivedStress	0	0.27	0	0
SCL.90R.Depression	0.5	0	0	0
SCL.90R.Somatization	0.49	0	0	0
SCL.90R.Anxiety	0.55	0	0	0
SCL.90R.Hostility	0.45	0	0	0
CSQ.Distraction	0	0	0	0.44
CSQ.IgnoringPain	0	0	0	0.48
CSQ.Distancing	0	0	0	0.47
CSQ.Coping	0	0	0	0.6
CSQ.Praying	0	0	0	0
STAIY1.StateTraitAnxiety	0	0.51	0	0
STAIY2.StateTraitAnxiety	0	0.53	0	0
PCS.Rumination	0	0	0.68	0
PCS.Magnification	0	0	0.43	0
PCS.Helplessness	0	0	0.59	0
EPQ.Escale	0	0	0	0
EPQ.Nscale	0	0	0	0
Percent Variance Explained	0.15	0.28	0.39	0.5

Table 5.13: Psychosocial Parameter Estimates and Different Thresholds, Component 1 (Chronic)

Method Details	Parameter Estimate	Standard Error	Odds Ratio	P-value
Traditional EFA/PCA	0.49992	0.05443		¡2e-16
SFA $c = 0.7$	0.35596	0.05258		1.29e-11
SFA $c = 0.6$	0.29673	0.03624		2.67e-16
SFA $c = 0.5$	0.41364	0.05211		2.05e-15
SFA $c = 0.4$	0.43485	0.05262		¡2e-16

Table 5.14: Psychosocial Parameter Estimates and Different Thresholds, Component 3 (Chronic)

Method Details	Parameter Estimate	Standard Error	Odds Ratio	P-value
Traditional EFA/PCA	0.23896	0.06946		0.000581
SFA $c = 0.7$	0.25626	0.06854		0.000185
SFA $c = 0.6$	0.24640	0.06832		0.00031
SFA $c = 0.5$	0.20356	0.06935		0.00333
SFA $c = 0.4$	0.17641	0.07025		0.012

Table 5.15: Psychosocial Parameter Estimates and Different Thresholds, Component 4 (Chronic)

Method Details	Parameter Estimate	Standard Error	Odds Ratio	P-value
Traditional EFA/PCA	-0.21275	0.07746		0.00602
SFA $c = 0.7$	-0.09480	0.07551		0.209
SFA $c = 0.6$	-0.16235	0.07676		0.0344
SFA $c = 0.5$	-0.17814	0.07708		0.0208
SFA $c = 0.4$	-0.18595	0.07722		0.016

CHAPTER 6: CONCLUSION

In this dissertation, three methods were developed to properly analyze data from the OPPERA study. Chapter 3 introduces a method for Cox modeling with missing censoring indicators, which allows for estimation of hazard ratios and incidence rates, even when a large subset of participants have uncertain event indicators. Event propabilities are modeled via logistic regression and outcomes are imputed repeatedly using those estimated probabilities. Implementing this method uncovered new risk factors of clinical importance. It also illustrated that incidence rates may be underestimated if the proposed method is not used. Incidence of TMD may be important for future studies in orofacial pain.

This method was well received by OPPERA investigators and widely applied, e.g. in (1). The methods paper based on Chapter 3 is in the process of submission. In its current state, the manuscript is appropriate for submission to a journal such as *Statistics in Medicine*. Additional work will be performed in hopes of proving consistency of the estimator. If this aim is successful, we will consider a more theoretical journal. Future work includes completing an extension that includes missing covariates (i.e. skipped screeners), missing time, and missing censoring indicators. This will allow study of individuals who have uncertain screener data as well as uncertain outcomes. In addition, there were a small number ($n < 20$) of false negatives, i.e. participants who screened negative but who actually had TMD. This number was too small of a sample size for meaningful modeling and imputation. Thus we did not make any further adjustments. In the future, the methodology could be extended to control for false negatives, or a

sensitivity analysis could be conducted.

Chapter 4 proposes a general method to analyze intermediate phenotypes in case-control studies. Subjects are weighted to correct for the oversampling of cases and undersampling of controls in this sample design. Standard methods are implemented with these weights to estimate the quantities of interest. Then bootstrapping is utilized to estimate the standard error. The method allows for analysis of outcomes that are currently difficult or impossible to study, such as time-to-event outcomes and sequencing statistics.

Additional work remains before the method may be published. First, application to haplotypes will be demonstrated. Second, the method will be applied to different outcomes in OPPERA. Third, the simulations on sequenced data will be completed. Fourth, the method may be applied to sequencing data from Glaxo-Smith-Klein (GSK). Once these steps are complete, the paper should be ready to submit to a genetic journal, such as *Genetic Epidemiology*.

Chapter 5 introduces a method for sparse factor analysis. Soft thresholding reduces the magnitude of each loading in a rigorous manner rather than arbitrarily ignoring loadings under a certain value. Sparse loadings result that are still predictive of the underlying outcome. Implications within the OPPERA study are wide. The sparse loadings in each domain may highlight which variables are most important for future studies and which variables may be dropped without losing significant information. It remains to discuss this work with the principal investigators. Additional simulations may be produced as well.

APPENDIX 1: SUPPLEMENTARY MATERIAL FOR CHAPTER 3

The primary objective of the OPPERA study is to identify possible risk factors for developing first-onset TMD. See Maixner et al. (75) and Slade et al. (113) for a more detailed description of the study. The risk factors considered in OPPERA are classified into the following domains: sociodemographic, clinical, psychosocial, autonomic, quantitative sensory testing (QST), and genetics. The remainder of this section describes these OPPERA domains in more detail.

First, sociodemographic information was recorded for each OPPERA participant. This includes age, gender, race, and OPPERA study site, as well as educational attainment, income, and marital status. For example, TMD is more common in females than males and in non-Hispanic whites than in other races. Details are provided in Slade et al. (113).

Clinical risk factors refer to variables that "typically are considered in clinical settings when evaluating patients" (86). These clinical variables may be evaluated via physical examinations or questionnaires. Examples include headaches, back aches, pain in other regions of the body, jaw mobility, jaw noises, and orofacial trauma. OPPERA participants also self-reported their health history, including the presence of comorbid pain conditions such as irritable bowel syndrome, fibromyalgia, and dysmenorrhea.

Psychosocial factors have also been shown to be associated with TMD (37). Specific qualities related to psychosocial functioning were evaluated in OPPERA, including general psychological function, affective distress, psychological stress, somatic awareness, and coping/catastrophizing. Affective distress measures include state and trait anxiety and mood. Psychological stress includes perceived stress and measures of post-traumatic stress disorder. Somatic awareness assesses sensitivity to physical sensations. Finally, coping/catastrophizing assesses individuals' ability to handle pain.

The association between TMD and the function of the autonomic nervous system was also evaluated. Key measures of autonomic function include blood pressure, heart rate, and heart rate variability, which were measured during the OPPERA baseline medical examination. In previous studies, TMD was associated with higher heart rates and lower heart rate variability, which are symptoms of dysregulation of the autonomic nervous system. See (76) for a more detailed description of the autonomic data collected in OPPERA.

The QST variables collected in OPPERA measure sensitivity to experimental pain. Several measures of experimental pain sensitivity were collected, including pressure pain thresholds measured by algometers, mechanical (pinprick) pain sensitivity, and thermal pain sensitivity. See (46) for a more detailed description of these QST variables.

Finally, the association between TMD and selected genetic markers was evaluated. A total of 3295 single nucleotide polymorphisms (SNP's) were selected from genes that are believed to be associated with pain. See (115) for more detail on how the SNP's were chosen and their association with TMD.

Overview of Additional Simulations

In this appendix, we provide the results of additional simulations. We investigate the performance of the method under a variety of missing data mechanisms. We also consider scenarios where the logistic regression model for estimating the probability of being a case is misspecified.

Recall that we created missing censoring indicators under the following classical missing data mechanisms of Rubin (1976):

(I) The probability of having a missing censoring indicator is independent of the data. This is known as missing completely at random (MCAR).

(II) The probability of having a missing censoring indicator depends on an observed

covariate. This is known as missing at random (MAR).

(III) The probability of having a missing censoring indicator depends on the censoring indicator itself. This is known as missing not at random (MNAR).

Simulations Under MCAR

When the data were MCAR, our method had less bias on average than the complete case method that depended on the true parameter value. Not only did our method have adequate coverage, but it had the most narrow confidence intervals of the methods with adequate coverage. As in other simulations, the method treating all missing indicators as failures had poor coverage and introduced extreme bias. The complete case method and the method that treat all missing censoring indicators as censored were valid, but had much wider confidence intervals than our method.

However, note that it would be dangerous to apply the complete case method to the OPPERA study. According to the OPPERA protocol, participants who do not screen positively and are not selected as matched controls will always be censored. Only participants who screen positively (i.e. those with $Q_i = 1$) should potentially have missing censoring indicators. Data that are MCAR allow participants with $Q_i = 0$ to have missing censoring indicators. Fitting the logistic regression model to those with $Q_i = 1$ only but generalizing to people with $Q_i = 0$ may result in extreme bias, as shown in section 6.

Additional Simulations Under MCAR

In order to more closely parallel the OPPERA study, we simulated data where we randomly set 40% of the censoring indicators to be missing for those with $Q_i = 1$. (Note that our simulations assume that censoring indicators can only be missing when $Q_i = 1$. Without this assumption the data would not be MCAR in this scenario, since

Q_i depends on X_{i2}, which is observed.) This setup assumes that the probability that a participant has a non-missing censoring indicator depends only on whether or not their screener was positive. The logistic regression model in this case included the covariates X_{i1} and X_{i2} as before, but not the time of the screener. Results are shown in Table A1.2. All methods had a negligible amount of bias in these scenarios except for the complete case method and the method that treated all missing indicators as failures. In these simulations, the complete case method also displayed extreme bias and poor coverage. This indicates that a complete case analysis would not be appropriate for a study such as OPPERA.

Alternative Logistic Regression Models

We considered several scenarios where the logistic regression model for the probability of being a case is misspecified. Recall that we originally modeled the probability of being a case as

$$P(\Delta_i = 1|X_i, \alpha) = \frac{\exp(\alpha' X_i + \gamma V_i)}{1 + \exp(\alpha' X_i + \gamma V_i)} \tag{A.1}$$

The original logistic model had the covariates $X_i = \{X_{i1}, X_{i2}\}$ and V_i where $X_{i1} \sim N(0, 2)$ and $X_{i2} \sim N(\Delta_i, 0.3)$ are mutually independent for $j = 1, 2, 3$ and $i = 1, \ldots, n$.

Two alternative models were examined:

1. The first alternative model was of the form (A.1) but used the covariates $\tilde{X}_i = \{X_{i1}, X_{i2}, X_{i3}\}$ and V_i where $X_{i3} \sim N(0, 1)$. This scenario was to used to evaluate the robustness of the method when an extraneous covariate is included in the model.

2. The second alternative model was generated according to (A.1) but was fit with the covariates $\tilde{X}_i = \{X_{i1}\}$ and V_i. In the context of OPPERA, this represents the scenario in which we failed to include a covariate that is associated with first-onset

TMD.

Tables A1.3 and A1.4 indicate that our method produces valid results even if a noisy variable is added to the model or if an important variable is not included in the model.

Next, we consider the scenario where censoring indicators may be missing even if a participant had a negative screener (i.e. $Q_i = 0$). For each such simulation, we randomly selected 40% of the observations to have missing censoring indicators regardless of the value of Q_i. In the first such simulation, the logistic regression model was correctly specified when $Q_i = 1$. (However, it will be applied to all observations with missing censoring indicators, including those for which $Q_i = 0$. Since the true value of the censoring indicator is always 0 when $Q_i = 0$, the model will be biased for these observations.) In the two remaining simulation scenarios, the model will be misspecified even when $Q_i = 1$ by either adding an extra covariate or leaving out a significant covariate as we did in the earlier simulations.

The results of these three additional simulations are shown in Tables A1.1, A1.6, and A1.7. The model performs well in two of the three scenarios, indicating that our methodology is robust against misspecification of the logistic regression model. However, when an important covariate is not included in the model, the estimates are badly biased. Empirical coverage ranged from 0% to 50%, significantly below the nominal rate. This indicates that our method can give incorrect results if the predictive accuracy of the logistic regression model is poor. Note that the method of Cook and Kosorok (8) also performs poorly in this scenario. If one cannot accurately estimate which censoring indicators are missing, it is unlikely that any method can produce valid confidence intervals for the Cox regression coefficients.

Simulations Under MNAR

We examined two possible scenarios where the data is MNAR:

(A) In the first, we set 30% of the censored observations and 50% of the failures to have missing indicators.

(B) In the second, we set 20% of the censored observations and 60% of the failures to have missing indicators.

Bias increased for all methods under both MNAR scenarios. In particular, the complete case method consistently displayed a high amount of bias and did not achieve the desired coverage rate. For our imputation method and the method of Cook and Kosorok (8), bias increased and coverage decreased as the true parameter value increased. This indicates that when the MAR assumption is violated, our method as well as the method of Cook and Kosorok (8) may not be valid. On the other hand, even when the data was not MAR, our method provided an improvement in terms of bias and coverage compared to the complete case method and the method that treats all missing subjects as failures. Moreover, the coverage probability was slightly greater for our method than for the method of Cook and Kosorok (8).

Simulations for Poisson Regression

We performed simulations to evaluate the performance of our method when the desired time-to-event analysis is a Poisson regression model rather than a Cox model. Poisson models are commonly used to estimate incidence rates, which is an objective of the OPPERA study.

The simulations were identical to those described in Section 3.4 except that the imputed data was used to fit Poisson regression models rather than Cox proportional hazards models. That is, we fit the data from imputations $j = 1, \ldots, m$ to the model

$$\log(\mu_i) = \alpha + \beta x_{i1} + \log(V_i). \tag{A.2}$$

where μ_i is the expected number of cases and the offset, $\log(V_i)$, is the logarithm of the survival time. We measured the bias, defined as $\hat{\beta}$ minus the true value, for $\beta \in \{-0.5, -1.5, -3\}$.

The Cook and Kosorok (8) method does not immediately generalize to Poisson regression. Consequently, we only compared our method to the unachievable ideal of no missing data, the complete case method, and the two ad hoc methods.

The use of Poisson regression allows us to estimate incidence rates. For each simulation, we estimated the incidence rate based on the coefficients of the Poisson regression model in (A.2). Specifically, estimated incidence rates for fixed values of X_{i1} are given by

$$\exp(\alpha + \beta x_{i1}) \tag{A.3}$$

The bias, confidence interval width, and coverage probability of each method are shown in Table A1.10. We also present the estimated incidence rates for each quartile of the random variable X_{i1} (i.e. the quartiles of the $N(2, 1)$ distribution). See Table A1.11.

Our method had close to 95% coverage probability when Poisson regression was used. None of the other methods had proper coverage for all of the simulations. Multiple imputation yielded the least bias of all the methods besides the unacheivable ideal of observing all data. It also produced more narrow confidence intervals than the complete case method and the method that treats all missing censoring indicators as censored.

The bias evident in parameter estimation was compounded for incidence rates. The complete case method and the two ad hoc methos consistently underestimated incidence. In fact, the complete case method underestimated incidence by about 30-200%. By contrast, our method differed from the unachievable ideal by only about 4-6%.

Table A1.1: Simulation Results for MCAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0010	0.0006	0.1671	0.0001	0.930
	Complete Case	-0.0042	0.0007	0.2218	0.0001	0.956
	Treat all as Censored	-0.0021	0.0007	0.2212	0.0001	0.947
	Treat all as Failures	0.0932	0.0005	0.1525	0.0001	0.323
	Cook & Kosorok	-0.0005	0.0006	0.1802	0.0002	0.943
	Multiple Imputation	-0.0005	0.0006	0.1713	0.0001	0.932
-1.5	Full Data	-0.0008	0.0010	0.3185	0.0002	0.966
	Complete Case	-0.0056	0.0014	0.4261	0.0003	0.948
	Treat all as Censored	-0.0025	0.0014	0.4229	0.0003	0.950
	Treat all as Failures	0.8277	0.0007	0.2036	0.0001	0.000
	Cook & Kosorok	-0.0003	0.0011	0.3512	0.0004	0.952
	Multiple Imputation	-0.0005	0.0011	0.3306	0.0002	0.947
-3	Full Data	-0.0190	0.0025	0.7574	0.0008	0.952
	Complete Case	-0.0321	0.0036	1.0255	0.0016	0.942
	Treat all as Censored	-0.0205	0.0035	1.0070	0.0015	0.950
	Treat all as Failures	2.5225	0.0009	0.2216	0.0001	0.000
	Cook & Kosorok	-0.0229	0.0028	0.9459	0.0024	0.953
	Multiple Imputation	-0.0239	0.0028	0.7961	0.0010	0.936

*: The Monte Carlo error is 0.007.

Table A1.2: Simulation Results for MCAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0010	0.0006	0.1671	0.0001	0.930
	Complete Case	-0.0543	0.0008	0.2210	0.0001	0.825
	Treat all as Censored	-0.0021	0.0007	0.2212	0.0001	0.947
	Treat all as Failures	0.0041	0.0006	0.1703	0.0001	0.929
	Cook & Kosorok	-0.0012	0.0006	0.1742	0.0001	0.934
	Multiple Imputation	-0.0012	0.0006	0.1722	0.0001	0.938
-1.5	Full Data	-0.0008	0.0010	0.3185	0.0002	0.966
	Complete Case	-0.1329	0.0014	0.4283	0.0003	0.759
	Treat all as Censored	-0.0025	0.0014	0.4229	0.0003	0.950
	Treat all as Failures	0.0849	0.0011	0.3147	0.0002	0.790
	Cook & Kosorok	-0.0006	0.0011	0.3411	0.0004	0.951
	Multiple Imputation	-0.0006	0.0011	0.3320	0.0002	0.957
-3	Full Data	-0.0190	0.0025	0.7574	0.0008	0.952
	Complete Case	-0.2342	0.0037	1.0336	0.0016	0.883
	Treat all as Censored	-0.0205	0.0035	1.0070	0.0015	0.950
	Treat all as Failures	0.6626	0.0025	0.6165	0.0006	0.047
	Cook & Kosorok	-0.0232	0.0028	0.8775	0.0016	0.940
	Multiple Imputation	-0.0240	0.0028	0.7996	0.0010	0.937

*: The Monte Carlo error is 0.007.

Table A1.3: Results for an Extra Covariate Included in the Logistic Regression Model

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0017	5e-04	0.1670	1e-04	0.957
	Complete Case	9e-04	7e-04	0.2156	1e-04	0.942
	Treat all as Censored	0.1038	7e-04	0.2132	1e-04	0.517
	Treat all as Failures	0.0010	6e-04	0.1703	1e-04	0.948
	Cook & Kosorok	-0.0016	6e-04	0.1741	1e-04	0.954
	Multiple Imputation	-0.0016	6e-04	0.1718	1e-04	0.942
-1.5	Full Data	-0.0010	0.001	0.3188	2e-04	0.950
	Complete Case	-0.0589	0.0014	0.4332	4e-04	0.929
	Treat all as Censored	0.1240	0.0014	0.4227	3e-04	0.765
	Treat all as Failures	0.0665	0.001	0.3166	2e-04	0.864
	Cook & Kosorok	-9e-04	0.0011	0.3425	3e-04	0.945
	Multiple Imputation	-0.0010	0.0011	0.3313	2e-04	0.946
-3	Full Data	-0.0170	0.0025	0.7586	8e-04	0.956
	Complete Case	-0.1892	0.0035	1.0783	0.0016	0.929
	Treat all as Censored	0.1090	0.0034	1.0351	0.0014	0.907
	Treat all as Failures	0.6010	0.0025	0.6275	6e-04	0.095
	Cook & Kosorok	-0.0184	0.0028	0.9093	0.0016	0.946
	Multiple Imputation	-0.0191	0.0028	0.8057	0.0012	0.941

*: The Monte Carlo error is 0.007.

Table A1.4: Results when a Relevant Covariate is Omitted frin the Logistic Regression Model

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0034	5e-04	0.1671	1e-04	0.950
	Complete Case	-0.0012	7e-04	0.2157	1e-04	0.943
	Treat all as Censored	0.1005	7e-04	0.2133	1e-04	0.544
	Treat all as Failures	-8e-04	6e-04	0.1704	1e-04	0.955
	Cook & Kosorok	-0.0036	6e-04	0.1739	1e-04	0.951
	Multiple Imputation	-0.0036	6e-04	0.1721	1e-04	0.954
-1.5	Full Data	-0.0055	0.001	0.3183	2e-04	0.955
	Complete Case	-0.0638	0.0014	0.4322	4e-04	0.925
	Treat all as Censored	0.1177	0.0014	0.4218	3e-04	0.787
	Treat all as Failures	0.0625	0.001	0.3161	2e-04	0.869
	Cook & Kosorok	-0.0054	0.0011	0.3451	4e-04	0.944
	Multiple Imputation	-0.0055	0.0011	0.3359	2e-04	0.954
-3	Full Data	-0.0190	0.0025	0.7574	8e-04	0.952
	Complete Case	-0.1888	0.0038	1.0804	0.0018	0.917
	Treat all as Censored	0.1134	0.0035	1.0355	0.0015	0.907
	Treat all as Failures	0.5996	0.0024	0.6264	6e-04	0.080
	Cook & Kosorok	-0.0165	0.0029	0.8985	0.0017	0.940
	Multiple Imputation	-0.0186	0.0029	0.8199	0.0011	0.943

*: The Monte Carlo error is 0.007.

Table A1.5: Results when the Logistic Regression Model is Applied to Observations with $Q_i = 0$

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0010	0.0006	0.1671	0.0001	0.930
	Complete Case	-0.0042	0.0007	0.2218	0.0001	0.956
	Treat all as Censored	-0.0021	0.0007	0.2212	0.0001	0.947
	Treat all as Failures	0.0932	0.0005	0.1525	0.0001	0.323
	Cook & Kosorok	-0.0005	0.0006	0.1802	0.0002	0.943
	Multiple Imputation	-0.0005	0.0006	0.1713	0.0001	0.932
-1.5	Full Data	-0.0008	0.0010	0.3185	0.0002	0.966
	Complete Case	-0.0056	0.0014	0.4261	0.0003	0.948
	Treat all as Censored	-0.0025	0.0014	0.4229	0.0003	0.950
	Treat all as Failures	0.8277	0.0007	0.2036	0.0001	0.000
	Cook & Kosorok	-0.0003	0.0011	0.3512	0.0004	0.952
	Multiple Imputation	-0.0005	0.0011	0.3306	0.0002	0.947
-3	Full Data	-0.0190	0.0025	0.7574	0.0008	0.952
	Complete Case	-0.0321	0.0036	1.0255	0.0016	0.942
	Treat all as Censored	-0.0205	0.0035	1.0070	0.0015	0.950
	Treat all as Failures	2.5225	0.0009	0.2216	0.0001	0.000
	Cook & Kosorok	-0.0229	0.0028	0.9459	0.0024	0.953
	Multiple Imputation	-0.0239	0.0028	0.7961	0.0010	0.936

*: The Monte Carlo error is 0.007.

Table A1.6: Results when an Extra Covariate is Included in the Logistic Regression Model and the Model is Applied to Observations with $Q_i = 0$

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0020	5e-04	0.1671	1e-04	0.957
	Complete Case	-0.0026	7e-04	0.2218	1e-04	0.954
	Treat all as Censored	-1e-04	7e-04	0.221	1e-04	0.945
	Treat all as Failures	0.0938	5e-04	0.1526	1e-04	0.328
	Cook & Kosorok	-4e-04	6e-04	0.1814	2e-04	0.951
	Multiple Imputation	-4e-04	6e-04	0.1713	1e-04	0.950
-1.5	Full Data	-0.0050	0.0011	0.3187	2e-04	0.943
	Complete Case	-0.0077	0.0014	0.4245	4e-04	0.949
	Treat all as Censored	-0.0038	0.0014	0.4218	3e-04	0.947
	Treat all as Failures	0.8225	7e-04	0.2040	1e-04	0.000
	Cook & Kosorok	-0.0041	0.0011	0.3526	4e-04	0.943
	Multiple Imputation	-0.0041	0.0011	0.3306	2e-04	0.939
-3	Full Data	-0.0191	0.0026	0.7586	0.0008	0.950
	Complete Case	-0.0389	0.0035	1.0301	0.0016	0.949
	Treat all as Censored	-0.0276	0.0034	1.0108	0.0014	0.947
	Treat all as Failures	2.5217	0.0008	0.2214	0.0001	0.000
	Cook & Kosorok	-0.0155	0.0029	0.9683	0.0025	0.961
	Multiple Imputation	-0.0156	0.0028	0.7978	0.0011	0.939

*: The Monte Carlo error is 0.007.

Table A1.7: Results when a Relevant Covariate is not Included in the Logistic Regression Model and the Model is Applied to Observations where $Q_i = 0$

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0001	0.0006	0.1667	0.0001	0.941
	Complete Case	0.0011	0.0008	0.2211	0.0001	0.939
	Treat all as Censored	0.0009	0.0007	0.2205	0.0001	0.939
	Treat all as Failures	0.0938	0.0005	0.1522	0.0001	0.331
	Cook & Kosorok	0.0840	0.0005	0.1581	0.0001	0.484
	Multiple Imputation	0.0840	0.0005	0.1551	0.0001	0.441
-1.5	Full Data	-0.0046	0.0011	0.3187	0.0002	0.941
	Complete Case	-0.0068	0.0014	0.4242	0.0004	0.939
	Treat all as Censored	-0.0029	0.0014	0.4217	0.0003	0.945
	Treat all as Failures	0.8236	0.0007	0.2041	0.0001	0.000
	Cook & Kosorok	0.5062	0.0013	0.3936	0.0006	0.003
	Multiple Imputation	0.5058	0.0013	0.2581	0.0002	0.000
-3	Full Data	-0.0229	0.0026	0.7606	0.0009	0.954
	Complete Case	-0.0335	0.0034	1.0296	0.0016	0.965
	Treat all as Censored	-0.0282	0.0034	1.0132	0.0014	0.956
	Treat all as Failures	2.5266	0.0008	0.2212	0.0001	0.000
	Cook & Kosorok	0.8890	0.0034	1.0710	0.0026	0.191
	Multiple Imputation	0.8867	0.0034	0.5424	0.0008	0.026

*: The Monte Carlo error is 0.007.

Table A1.8: Simulation Results for MNAR, scenario A

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0021	0.0006	0.1671	0.0001	0.938
	Complete Case	-0.0755	0.0008	0.2424	0.0004	0.778
	Treat all as Censored	-0.0022	0.0008	0.2421	0.0001	0.942
	Treat all as Failures	0.0023	0.0006	0.1705	0.0001	0.942
	Cook & Kosorok	-0.0054	0.0006	0.1760	0.0001	0.940
	Multiple Imputation	-0.0054	0.0006	0.1732	0.0001	0.943
-1.5	Full Data	-0.0030	0.0011	0.3185	0.0002	0.942
	Complete Case	-0.1744	0.0016	0.4691	0.0004	0.717
	Treat all as Censored	-0.0049	0.0015	0.4623	0.0004	0.947
	Treat all as Failures	0.0646	0.0011	0.3172	0.0002	0.875
	Cook & Kosorok	-0.0206	0.0011	0.3460	0.0004	0.921
	Multiple Imputation	-0.0207	0.0011	0.3362	0.0002	0.939
-3	Full Data	-0.0289	0.0026	0.7611	0.0009	0.948
	Complete Case	-0.3308	0.0040	1.1490	0.0018	0.824
	Treat all as Censored	-0.0339	0.0039	1.1060	0.0016	0.935
	Treat all as Failures	0.5242	0.0026	0.6487	0.0007	0.181
	Cook & Kosorok	-0.0737	0.0029	0.9026	0.0018	0.898
	Multiple Imputation	-0.0745	0.0029	0.8194	0.0011	0.940

*: The Monte Carlo error is 0.007.

Table A1.9: Simulation Results for MNAR, scenario B

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0021	0.0006	0.1671	0.0001	0.938
	Complete Case	-0.0998	0.0009	0.2713	0.0002	0.687
	Treat all as Censored	-0.0016	0.0009	0.2707	0.0001	0.943
	Treat all as Failures	0.0009	0.0006	0.1707	0.0001	0.943
	Cook & Kosorok	-0.0094	0.0006	0.1783	0.0001	0.933
	Multiple Imputation	-0.0095	0.0006	0.1747	0.0001	0.934
-1.5	Full Data	-0.0030	0.0011	0.3185	0.0002	0.942
	Complete Case	-0.2278	0.0018	0.5289	0.0005	0.618
	Treat all as Censored	-0.0046	0.0017	0.5169	0.0004	0.958
	Treat all as Failures	0.0444	0.0011	0.3201	0.0002	0.910
	Cook & Kosorok	-0.0398	0.0012	0.3534	0.0004	0.902
	Multiple Imputation	-0.0398	0.0012	0.3417	0.0002	0.920
-3	Full Data	-0.0289	0.0026	0.7611	0.0009	0.948
	Complete Case	-0.4316	0.0046	1.3085	0.0023	0.781
	Treat all as Censored	-0.0464	0.0044	1.2434	0.002	0.932
	Treat all as Failures	0.3661	0.0027	0.6850	0.0007	0.451
	Cook & Kosorok	-0.1180	0.0030	0.9425	0.0021	0.870
	Multiple Imputation	-0.1189	0.0030	0.8394	0.0012	0.924

*: The Monte Carlo error is 0.007.

Table A1.10: Simulation Results for Poisson Models, MAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0036	0.0005	0.1596	0.0001	0.956
	Complete Case	-0.0101	0.0007	0.2067	0.0001	0.948
	Treat all as Censored	0.0813	0.0007	0.2049	0.0001	0.652
	Treat all as Failures	-0.0002	0.0006	0.1628	0.0001	0.957
	Multiple Imputation	-0.0036	0.0006	0.1642	0.0001	0.959
-1.5	Full Data	-0.0005	0.0010	0.2864	0.0002	0.945
	Complete Case	-0.1008	0.0015	0.3956	0.0004	0.829
	Treat all as Censored	0.0704	0.0014	0.3883	0.0003	0.898
	Treat all as Failures	0.0717	0.0010	0.2857	0.0002	0.820
	Multiple Imputation	0.0006	0.0011	0.2979	0.0002	0.940
-3	Full Data	-0.0184	0.0021	0.5994	0.0007	0.958
	Complete Case	-0.2733	0.0032	0.8710	0.0016	0.792
	Treat all as Censored	0.0206	0.0030	0.8485	0.0012	0.954
	Treat all as Failures	0.5232	0.0025	0.5544	0.0006	0.085
	Multiple Imputation	-0.0219	0.0024	0.6353	0.0009	0.940

*: The Monte Carlo error is 0.007.

Table A1.11: Simulation Results for Incidence Rates

β	Method	Q1	SE	Q2	SE	Q3	SE
-0.5	Full Data	0.5157	0.0003	0.3671	0.0002	0.2615	0.0002
	Complete Case	0.4261	0.0004	0.3019	0.0002	0.2142	0.0002
	Treat all as Censored	0.2991	0.0003	0.2254	0.0002	0.1701	0.0002
	Treat all as Failures	0.4949	0.0003	0.3531	0.0002	0.2521	0.0002
	Multiple Imputation	0.4910	0.0003	0.3495	0.0002	0.2490	0.0002
-1.5	Full Data	0.1375	0.0001	0.0501	0.0001	0.0183	0.0000
	Complete Case	0.0968	0.0001	0.0330	0.0001	0.0113	0.0000
	Treat all as Censored	0.0753	0.0001	0.0288	0.0000	0.0110	0.0000
	Treat all as Failures	0.1388	0.0001	0.0530	0.0001	0.0203	0.0000
	Multiple Imputation	0.1309	0.0001	0.0477	0.0001	0.0174	0.0000
-3	Full Data	0.0186	0.0000	0.0025	0.0000	0.0003	0.0000
	Complete Case	0.0106	0.0000	0.0012	0.0000	0.0001	0.0000
	Treat all as Censored	0.0097	0.0000	0.0013	0.0000	0.0002	0.0000
	Treat all as Failures	0.0301	0.0001	0.0058	0.0000	0.0011	0.0000
	Multiple Imputation	0.0176	0.0000	0.0023	0.0000	0.0003	0.0000

*: Q1 denotes rates based on the lower quartile of X_{i1}.

*: Q2 denotes rates based on the median of X_{i1}.

*: Q3 denotes rates based on the upper quartile of X_{i1}.

APPENDIX 2: SUPPLEMENTARY MATERIAL FOR CHAPTER 5

Results for the QST domain remained unchanged in overall interpretation for different threshold values, ranging from 0.4 to 0.8 for all components. For the full cohort of 3443 individuals, see Tables A2.1-A2.4 and Table 5.9. For the follow-up cohort of 2737 individuals see Tables A2.6-A2.9. As expected, results are very close for the full OPPERA dataset compared to for those with follow-up data only. No further discussion is warranted, but tables are provided for the reader.

Table A2.1: PCA/SFA QST Results for the Entire OPPERA Cohort, c=0.4

Component	1	2	3	4	5
Pressure Pain Threshold					
Temporalis	0	0	0.42	0	0
Masseter	0	0	0.44	0	0
TMJ	0	0	0.42	0	0
Trapezius	0	0	0.41	0	0
Epicondyl	0	0	0.37	0	0
Mechanical Cutaneous Pain Threshhold	0	0	0.38	0	0
Single Stimulus					
256 mN Probe	0	0	0	0.36	0
512 mN Probe	0	0	0	0.37	0
After Sensation					
15 s, 256 mN probe	0	0	0	0.48	0
30 s, 256 mN probe	0	0	0	0.40	0
15 s, 512 mN probe	0	0.04	0	0.41	0
30 s, 512 mN probe	0	0.07	0	0.37	0
Temporal Summation					
256 mN probe	0	0	0	0.19	0
512 mN probe	0	0	0	0	0
Heat Pain Tolerance	-0.03	0	0	0	0
Single Stimulus Ratings					
46	0.35	0	0	0	0
48	0.42	0	0	0	0
50	0.43	0	0	0	0
Area Under the Curve					
46	0.42	0	0	0	0
48	0.45	0	0	0	0
50	0.39	0	0	0	0
After Sensations					
15 s, 46	0	0.37	0	0	0
30 s, 46	0	0.40	0	0	0
15 s, 48	0	0.41	0	0	0
30 s, 48	0	0.44	0	0	0
15 s, 50	0	0.40	0	0	0
30 s, 50	0	0.42	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.46
48	0	0	0	0	0.50
50	0	0	0	0	0.34
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0.28
48	0	0	0	0	0.43
50	0	0	0	0	0.39
Percent Variance Explained	0.15	0.31	0.45	0.58	0.69

Table A2.2: PCA/SFA QST Results for the Entire OPPERA Cohort, c=0.5

Component	1	2	3	4	5
Pressure Pain Threshold					
Temporalis	0	0	0.42	0	0
Masseter	0	0	0.45	0	0
TMJ	0	0	0.43	0	0
Trapezius	0	0	0.41	0	0
Epicondyl	0	0	0.36	0	0
Mechanical Cutaneous Pain Threshold	0	0	0.37	0	0
Single Stimulus					
256 mN Probe	0	0	0	0.34	0
512 mN Probe	0	0	0	0.37	0
Temporal Summation					
256 mN probe	0	0	0	0.1	0
512 mN probe	0	0	0	0	0
After Sensation Ratings					
15 s, 256 mN probe	0	0	0	0.52	0
15 s, 512 mN probe	0	0	0	0.42	0
30 s, 256 mN probe	0	0	0	0.4	0
30 s, 512 mN probe	0	0	0	0.35	0
Heat Pain Tolerance	0	0	0	0	0
After Sensations					
15 s, 46	0	0.36	0	0	0
15 s, 48	0	0.41	0	0	0
15 s, 50	0	0.4	0	0	0
30 s, 46	0	0.4	0	0	0
30 s, 48	0	0.45	0	0	0
30 s, 50	0	0.43	0	0	0
Single Stimulus					
46	0.33	0	0	0	0
48	0.42	0	0	0	0
50	0.43	0	0	0	0
Area Under the Curve					
46	0.42	0	0	0	0
48	0.46	0	0	0	0
50	0.37	0	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.47
48	0	0	0	0	0.53
50	0	0	0	0	0.32
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0.24
48	0	0	0	0	0.44
50	0	0	0	0	0.38
Percent Variance Explained	0.15	0.3	0.44	0.57	0.68

Table A2.3: PCA/SFA QST Results for the Entire OPPERA Cohort, c=0.6

Component	1	2	3	4	5
Pressure Pain Threshold					
Temporalis	0	0	0.43	0	0
Masseter	0	0	0.47	0	0
TMJ	0	0	0.44	0	0
Trapezius	0	0	0.41	0	0
Epicondyl	0	0	0.34	0	0
Mechanical Cutaneous Pain Threshold	0	0	0.36	0	0
Single Stimulus					
256 mN Probe	0	0	0	0.3	0
512 mN Probe	0	0	0	0.34	0
Temporal Summation					
256 mN probe	0	0	0	0	0
512 mN probe	0	0	0	0	0
After Sensation Ratings					
15 s, 256 mN probe	0	0	0	0.59	0
15 s, 512 mN probe	0	0	0	0.43	0
30 s, 256 mN probe	0	0	0	0.4	0
30 s, 512 mN probe	0	0	0	0.32	0
Heat Pain Tolerance	0	0	0	0	0
After Sensations					
15 s, 46	0	0.34	0	0	0
15 s, 48	0	0.41	0	0	0
15 s, 50	0	0.39	0	0	0
30 s, 46	0	0.39	0	0	0
30 s, 48	0	0.46	0	0	0
30 s, 50	0	0.44	0	0	0
Single Stimulus					
46	0.3	0	0	0	0
48	0.42	0	0	0	0
50	0.44	0	0	0	0
Area Under the Curve					
46	0.43	0	0	0	0
48	0.48	0	0	0	0
50	0.36	0	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.5
48	0	0	0	0	0.59
50	0	0	0	0	0.27
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0.15
48	0	0	0	0	0.44
50	0	0	0	0	0.35
Percent Variance Explained	0.15	0.30	0.44	0.56	0.66

Table A2.4: PCA/SFA QST Results for the Entire OPPERA Cohort, c=0.7

Component	1	2	3	4	5
Pressure Pain Threshold					
Temporalis	0	0	0.43	0	0
Masseter	0	0	0.5	0	0
TMJ	0	0	0.45	0	0
Trapezius	0	0	0.41	0	0
Epicondyl	0	0	0.29	0	0
Mechanical Cutaneous Pain Threshold	0	0	0.33	0	0
Single Stimulus					
256 mN Probe	0	0	0	0.12	0
512 mN Probe	0	0	0	0.21	0
Temporal Summation					
256 mN probe	0	0	0	0	0
512 mN probe	0	0	0	0	0
After Sensation Ratings					
15 s, 256 mN probe	0	0	0	0.79	0
15 s, 512 mN probe	0	0	0	0.42	0
30 s, 256 mN probe	0	0	0	0.34	0
30 s, 512 mN probe	0	0	0	0.16	0
Heat Pain Tolerance	0	0	0	0	0
After Sensations					
15 s, 46	0	0.3	0	0	0
15 s, 48	0	0.41	0	0	0
15 s, 50	0	0.38	0	0	0
30 s, 46	0	0.38	0	0	0
30 s, 48	0	0.5	0	0	0
30 s, 50	0	0.46	0	0	0
Single Stimulus					
46	0.22	0	0	0	0
48	0.42	0	0	0	0
50	0.45	0	0	0	0
Area Under the Curve					
46	0.44	0	0	0	0
48	0.52	0	0	0	0
50	0.33	0	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.52
48	0	0	0	0	0.69
50	0	0	0	0	0.09
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0
48	0	0	0	0	0.42
50	0	0	0	0	0.25
Percent Variance Explained	0.15	0.29	0.43	0.53	0.61

Table A2.5: PCA/SFA QST Results for the OPPERA Follow-up Cohort, c=0.4

Component	1	2	3	4	5
Pressure Pain Threshold					
Temporalis	0	0	0.42	0	0
Masseter	0	0	0.44	0	0
TMJ	0	0	0.42	0	0
Trapezius	0	0	0.41	0	0
Epicondyl	0	0	0.36	0	0
Mechanical Cutaneous Pain Threshold	0	0	0.39	0	0
Single Stimulus					
256 mN Probe	0	0	0	0.36	0
512 mN Probe	0	0	0	0.37	0
Temporal Summation					
256 mN probe	0	0	0	0.19	0
512 mN probe	0	0	0	0	0
After Sensation Ratings					
15 s, 256 mN probe	0	0	0	0.48	0
15 s, 512 mN probe	0	0.06	0	0.41	0
30 s, 256 mN probe	0	0	0	0.4	0
30 s, 512 mN probe	0	0.09	0	0.35	0
Heat Pain Tolerance	-0.03	0	0	0	0
After Sensations					
15 s, 46	0	0.37	0	0	0
15 s, 48	0	0.41	0	0	0
15 s, 50	0	0.4	0	0	0
30 s, 46	0	0.4	0	0	0
30 s, 48	0	0.43	0	0	0
30 s, 50	0	0.42	0	0	0
Single Stimulus					
46	0.35	0	0	0	0
48	0.42	0	0	0	0
50	0.43	0	0	0	0
Area Under the Curve					
46	0.42	0	0	0	0
48	0.44	0	0	0	0
50	0.38	0	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.46
48	0	0	0	0	0.51
50	0	0	0	0	0.35
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0.27
48	0	0	0	0	0.43
50	0	0	0	0	0.39
Percent Variance Explained	0.15	0.31	0.45	0.58	0.69

Table A2.6: PCA/SFA QST Results for the OPPERA Follow-up Cohort, c=0.5

Component	1	2	3	4	5
Pressure Pain Threshold					
Temporalis	0	0	0.42	0	0
Masseter	0	0	0.45	0	0
TMJ	0	0	0.43	0	0
Trapezius	0	0	0.41	0	0
Epicondyl	0	0	0.35	0	0
Mechanical Cutaneous Pain Threshold	0	0	0.38	0	0
Single Stimulus					
256 mN Probe	0	0	0	0.35	0
512 mN Probe	0	0	0	0.37	0
Temporal Summation					
256 mN probe	0	0	0	0.11	0
512 mN probe	0	0	0	0	0
After Sensation Ratings					
15 s, 256 mN probe	0	0	0	0.52	0
15 s, 512 mN probe	0	0	0	0.42	0
30 s, 256 mN probe	0	0	0	0.41	0
30 s, 512 mN probe	0	0	0	0.34	0
Heat Pain Tolerance	0	0	0	0	0
After Sensations					
15 s, 46	0	0.36	0	0	0
15 s, 48	0	0.41	0	0	0
15 s, 50	0	0.4	0	0	0
30 s, 46	0	0.4	0	0	0
30 s, 48	0	0.44	0	0	0
30 s, 50	0	0.43	0	0	0
Single Stimulus					
46	0.33	0	0	0	0
48	0.42	0	0	0	0
50	0.43	0	0	0	0
Area Under the Curve					
46	0.42	0	0	0	0
48	0.45	0	0	0	0
50	0.37	0	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.47
48	0	0	0	0	0.54
50	0	0	0	0	0.33
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0.22
48	0	0	0	0	0.43
50	0	0	0	0	0.38
Percent Variance Explained	0.15	0.3	0.44	0.57	0.68

Table A2.7: PCA/SFA QST Results for the OPPERA Follow-up Cohort, c=0.6

Component	1	2	3	4	5
Pressure Pain Threshold					
Temporalis	0	0	0.42	0	0
Masseter	0	0	0.46	0	0
TMJ	0	0	0.43	0	0
Trapezius	0	0	0.41	0	0
Epicondyl	0	0	0.33	0	0
Mechanical Cutaneous Pain Threshold	0	0	0.38	0	0
Single Stimulus					
256 mN Probe	0	0	0	0.31	0
512 mN Probe	0	0	0	0.34	0
Temporal Summation					
256 mN probe	0	0	0	0	0
512 mN probe	0	0	0	0	0
After Sensation Ratings					
15 s, 256 mN probe	0	0	0	0.59	0
15 s, 512 mN probe	0	0	0	0.43	0
30 s, 256 mN probe	0	0	0	0.41	0
30 s, 512 mN probe	0	0	0	0.3	0
Heat Pain Tolerance	0	0	0	0	0
After Sensations					
15 s, 46	0	0.34	0	0	0
15 s, 48	0	0.41	0	0	0
15 s, 50	0	0.4	0	0	0
30 s, 46	0	0.39	0	0	0
30 s, 48	0	0.46	0	0	0
30 s, 50	0	0.44	0	0	0
Single Stimulus					
46	0.3	0	0	0	0
48	0.43	0	0	0	0
50	0.44	0	0	0	0
Area Under the Curve					
46	0.43	0	0	0	0
48	0.47	0	0	0	0
50	0.35	0	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.49
48	0	0	0	0	0.59
50	0	0	0	0	0.28
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0.12
48	0	0	0	0	0.44
50	0	0	0	0	0.36
Percent Variance Explained	0.15	0.3	0.44	0.56	0.66

Table A2.8: PCA/SFA QST Results for the OPPERA Follow-up Cohort, c=0.7

Component	1	2	3	4	5
Pressure Pain Threshold					
Temporalis	0	0	0.43	0	0
Masseter	0	0	0.5	0	0
TMJ	0	0	0.44	0	0
Trapezius	0	0	0.41	0	0
Epicondyl	0	0	0.28	0	0
Mechanical Cutaneous Pain Threshold	0	0	0.35	0	0
Single Stimulus					
256 mN Probe	0	0	0	0.12	0
512 mN Probe	0	0	0	0.2	0
Temporal Summation					
256 mN probe	0	0	0	0	0
512 mN probe	0	0	0	0	0
After Sensation Ratings					
15 s, 256 mN probe	0	0	0	0.8	0
15 s, 512 mN probe	0	0	0	0.41	0
30 s, 256 mN probe	0	0	0	0.35	0
30 s, 512 mN probe	0	0	0	0.09	0
Heat Pain Tolerance	0	0	0	0	0
After Sensations					
15 s, 46	0	0.29	0	0	0
15 s, 48	0	0.41	0	0	0
15 s, 50	0	0.39	0	0	0
30 s, 46	0	0.38	0	0	0
30 s, 48	0	0.49	0	0	0
30 s, 50	0	0.46	0	0	0
Single Stimulus					
46	0.23	0	0	0	0
48	0.44	0	0	0	0
50	0.46	0	0	0	0
Area Under the Curve					
46	0.43	0	0	0	0
48	0.5	0	0	0	0
50	0.32	0	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.51
48	0	0	0	0	0.7
50	0	0	0	0	0.11
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0
48	0	0	0	0	0.41
50	0	0	0	0	0.26
Percent Variance Explained	0.15	0.29	0.43	0.52	0.61

Table A2.9: PCA/SFA QST Results for the OPPERA Follow-up Cohort, c=0.8

Component	1	2	3	4	5
Pressure Pain Threshold					
Temporalis	0	0	0.44	0	0
Masseter	0	0	0.6	0	0
TMJ	0	0	0.47	0	0
Trapezius	0	0	0.39	0	0
Epicondyl	0	0	0.09	0	0
Mechanical Cutaneous Pain Threshold	0	0	0.26	0	0
Single Stimulus					
256 mN Probe	0	0	0	0	0
512 mN Probe	0	0	0	0	0
Temporal Summation					
256 mN probe	0	0	0	0	0
512 mN probe	0	0	0	0	0
After Sensation Ratings					
15 s, 256 mN probe	0	0	0	1	0
15 s, 512 mN probe	0	0	0	0	0
30 s, 256 mN probe	0	0	0	0	0
30 s, 512 mN probe	0	0	0	0	0
Heat Pain Tolerance	0	0	0	0	0
After Sensations					
15 s, 46	0	0.06	0	0	0
15 s, 48	0	0.37	0	0	0
15 s, 50	0	0.34	0	0	0
30 s, 46	0	0.31	0	0	0
30 s, 48	0	0.61	0	0	0
30 s, 50	0	0.52	0	0	0
Single Stimulus					
46	0	0	0	0	0
0 48	0.44	0	0	0	0
50	0.49	0	0	0	0
Area Under the Curve					
46	0.43	0	0	0	0
48	0.6	0	0	0	0
50	0.14	0	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0	0	0.32
48	0	0	0	0	0.95
50	0	0	0	0	0
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0	0	0
48	0	0	0	0	0
50	0	0	0	0	0
Percent Variance Explained	0.12	0.24	0.37	0.4	0.44

Table A2.10: PCA/SFA Autonomic Results for the OPPERA Followup Cohort, c=0.4

Component	1	2	3	4	5
AvgRestingSPB	0.29	0	0	0	0
AvgRestingDPB	0.22	0	0	0	0
AvgRestingMAP	0.29	0	0	0	0
AvgRestingHR	0	0	0.3	0	0
InitialOrthoSPB	0.18	0	0	0	0
InitialOrthoDPB	0.2	0	0	0	0
InitialOrthoMAP	0.26	0	0	0	0
InitialOrthoHR	0	0	0.2	0	0
StroopColorSPBmean	0.33	0	0	0	0
StroopColorDPBmean	0.31	0	0	0	0
StroopColorMAPmean	0.35	0	0	0	0
StroopColorHRmean	0	0	0.4	0	0
StroopEmotionalSPBmean	0.33	0	0	0	0
StroopEmotionalDPBmean	0.29	0	0	0	0
StroopEmotionalMAPmean	0.35	0	0	0	0
StroopEmotionalHRmean	0	0	0.41	0	0
QST.SHRVBaseline20MinMeanHR	0	0	0.3	0	0
QST.SHRVBaseline20MinSDNN	0	0	0	0.34	0
QST.SHRVBaseline20MinRMSSD	0	0	0	0.12	0
HRVBaseline20minLnTP	0	0	0	0.53	0
HRVBaseline20minLnVLF	0	0	0	0.41	0
HRVBaseline20minLnLF	0	0	0	0.49	0
HRVBaseline20minLnHF	0	0	0	0.42	0
QST.SHRVOrthostatic5MinMeanHR	0	0	0.33	0	0
QST.SHRVOrthostatic5MinSDNN	0	0	0	0	0.47
QST.SHRVOrthostatic5MinRMSSD	0	0	0	0	0.38
HRVOrtho5minLnTP	0	0	0	0	0.45
HRVOrtho5minLnVLF	0	0	0	0	0.33
HRVOrtho5minLnLF	0	0	0	0	0.38
HRVOrtho5minLnHF	0	0	0	0	0.43
QST.SHRVSTROOPColor5MinMeanHR	0	0	0.41	0	0
QST.SHRVSTROOPColor5MinSDNN	0	0.42	0	0	0
QST.SHRVSTROOPColor5MinRMSSD	0	0.38	0	0	0
HRVStroopColor5minLnTP	0	0.32	0	0	0
HRVStroopColor5minLnVLF	0	0.18	0	0	0
HRVStroopColor5minLnLF	0	0.28	0	0	0
HRVStroopColor5minLnHF	0	0.33	0	0	0
QST.SHRVSTROOPEmotion5mnMeanHR	0	0	0.42	0	0
QST.SHRVSTROOPEmotion5mnSDNN	0	0.38	0	0	0
QST.SHRVSTROOPEmotion5mnRMSSD	0	0.36	0	0	0
HRVStroopEmotion5minLnTP	0	0.26	0	0	0
HRVStroopEmotion5minLnVLF	0	0.13	0	0	0
Percent Variance Explained	0.2	0.36	0.51	0.61	0.72

Table A2.11: PCA/SFA Autonomic Results for the OPPERA Followup Cohort, c=0.5

Component	1	2	3	4	5
AvgRestingSPB	0.29	0	0	0	0
AvgRestingDPB	0.2	0	0	0	0
AvgRestingMAP	0.29	0	0	0	0
AvgRestingHR	0	0	0.28	0	0
InitialOrthoSPB	0.15	0	0	0	0
InitialOrthoDPB	0.18	0	0	0	0
InitialOrthoMAP	0.25	0	0	0	0
InitialOrthoHR	0	0	0.15	0	0
StroopColorSPBmean	0.34	0	0	0	0
StroopColorDPBmean	0.31	0	0	0	0
StroopColorMAPmean	0.37	0	0	0	0
StroopColorHRmean	0	0	0.41	0	0
StroopEmotionalSPBmean	0.33	0	0	0	0
StroopEmotionalDPBmean	0.29	0	0	0	0
StroopEmotionalMAPmean	0.36	0	0	0	0
StroopEmotionalHRmean	0	0	0.42	0	0
QST.SHRVBaseline20MinMeanHR	0	0	0.28	0	0
QST.SHRVBaseline20MinSDNN	0	0	0	0.3	0
QST.SHRVBaseline20MinRMSSD	0	0	0	0.01	0
HRVBaseline20minLnTP	0	0	0	0.57	0
HRVBaseline20minLnVLF	0	0	0	0.39	0
HRVBaseline20minLnLF	0	0	0	0.51	0
HRVBaseline20minLnHF	0	0	0	0.42	0
QST.SHRVOrthostatic5MinMeanHR	0	0	0.32	0	0
QST.SHRVOrthostatic5MinSDNN	0	0	0	0	0.48
QST.SHRVOrthostatic5MinRMSSD	0	0	0	0	0.37
HRVOrtho5minLnTP	0	0	0	0	0.46
HRVOrtho5minLnVLF	0	0	0	0	0.29
HRVOrtho5minLnLF	0	0	0	0	0.37
HRVOrtho5minLnHF	0	0	0	0	0.44
QST.SHRVSTROOPColor5MinMeanHR	0	0	0.42	0	0
QST.SHRVSTROOPColor5MinSDNN	0	0.44	0	0	0
QST.SHRVSTROOPColor5MinRMSSD	0	0.39	0	0	0
HRVStroopColor5minLnTP	0	0.32	0	0	0
HRVStroopColor5minLnVLF	0	0.13	0	0	0
HRVStroopColor5minLnLF	0	0.27	0	0	0
HRVStroopColor5minLnHF	0	0.33	0	0	0
QST.SHRVSTROOPEmotion5mnMeanHR	0	0	0.44	0	0
QST.SHRVSTROOPEmotion5mnSDNN	0	0.39	0	0	0
QST.SHRVSTROOPEmotion5mnRMSSD	0	0.37	0	0	0
HRVStroopEmotion5minLnTP	0	0.24	0	0	0
HRVStroopEmotion5minLnVLF	0	0.07	0	0	0
Percent Variance Explained	0.2	0.36	0.5	0.59	0.7

Table A2.12: PCA/SFA Autonomic Results for the OPPERA Followup Cohort, c=0.6

Component	1	2	3	4	5
AvgRestingSPB	0.29	0	0	0	0
AvgRestingDPB	0.16	0	0	0	0
AvgRestingMAP	0.29	0	0	0	0
AvgRestingHR	0	0	0.25	0	0
InitialOrthoSPB	0.09	0	0	0	0
InitialOrthoDPB	0.13	0	0	0	0
InitialOrthoMAP	0.23	0	0	0	0
InitialOrthoHR	0	0	0.06	0	0
StroopColorSPBmean	0.35	0	0	0	0
StroopColorDPBmean	0.31	0	0	0	0
StroopColorMAPmean	0.4	0	0	0	0
StroopColorHRmean	0	0	0.42	0	0
StroopEmotionalSPBmean	0.35	0	0	0	0
StroopEmotionalDPBmean	0.29	0	0	0	0
StroopEmotionalMAPmean	0.39	0	0	0	0
StroopEmotionalHRmean	0	0	0.44	0	0
QST.SHRVBaseline20MinMeanHR	0	0	0.24	0	0
QST.SHRVBaseline20MinSDNN	0	0	0	0.23	0
QST.SHRVBaseline20MinRMSSD	0	0	0	0	0
HRVBaseline20minLnTP	0	0	0	0.61	0
HRVBaseline20minLnVLF	0	0	0	0.36	0
HRVBaseline20minLnLF	0	0	0	0.53	0
HRVBaseline20minLnHF	0	0	0	0.4	0
QST.SHRVOrthostatic5MinMeanHR	0	0	0.3	0	0
QST.SHRVOrthostatic5MinSDNN	0	0	0	0	0.52
QST.SHRVOrthostatic5MinRMSSD	0	0	0	0	0.35
HRVOrtho5minLnTP	0	0	0	0	0.49
HRVOrtho5minLnVLF	0	0	0	0	0.23
HRVOrtho5minLnLF	0	0	0	0	0.34
HRVOrtho5minLnHF	0	0	0	0	0.45
QST.SHRVSTROOPColor5MinMeanHR	0	0	0.44	0	0
QST.SHRVSTROOPColor5MinSDNN	0	0.48	0	0	0
QST.SHRVSTROOPColor5MinRMSSD	0	0.41	0	0	0
HRVStroopColor5minLnTP	0	0.31	0	0	0
HRVStroopColor5minLnVLF	0	0.02	0	0	0
HRVStroopColor5minLnLF	0	0.23	0	0	0
HRVStroopColor5minLnHF	0	0.32	0	0	0
QST.SHRVSTROOPEmotion5mnMeanHR	0	0	0.46	0	0
QST.SHRVSTROOPEmotion5mnSDNN	0	0.41	0	0	0
QST.SHRVSTROOPEmotion5mnRMSSD	0	0.38	0	0	0
HRVStroopEmotion5minLnTP	0	0.2	0	0	0
HRVStroopEmotion5minLnVLF	0	0	0	0	0
Percent Variance Explained	0.19	0.33	0.47	0.56	0.67

Table A2.13: PCA/SFA Autonomic Results for the OPPERA Followup Cohort, c=0.8

Component	1	2	3	4	5
AvgRestingSPB	0.17	0	0	0	0
AvgRestingDPB	0	0	0	0	0
AvgRestingMAP	0.18	0	0	0	0
AvgRestingHR	0	0	0	0	0
InitialOrthoSPB	0	0	0	0	0
InitialOrthoDPB	0	0	0	0	0
InitialOrthoMAP	0	0	0	0	0
InitialOrthoHR	0	0	0	0	0
StroopColorSPBmean	0.39	0	0	0	0
StroopColorDPBmean	0.27	0	0	0	0
StroopColorMAPmean	0.53	0	0	0	0
StroopColorHRmean	0	0	0.43	0	0
StroopEmotionalSPBmean	0.37	0	0	0	0
StroopEmotionalDPBmean	0.18	0	0	0	0
StroopEmotionalMAPmean	0.51	0	0	0	0
StroopEmotionalHRmean	0	0	0.5	0	0
QST.SHRVBaseline20MinMeanHR	0	0	0	0	0
QST.SHRVBaseline20MinSDNN	0	0	0	0	0
QST.SHRVBaseline20MinRMSSD	0	0	0	0	0
HRVBaseline20minLnTP	0	0	0	0.86	0
HRVBaseline20minLnVLF	0	0	0	0	0
HRVBaseline20minLnLF	0	0	0	0.51	0
HRVBaseline20minLnHF	0	0	0	0	0
QST.SHRVOrthostatic5MinMeanHR	0	0	0	0	0
QST.SHRVOrthostatic5MinSDNN	0	0	0	0	0.79
QST.SHRVOrthostatic5MinRMSSD	0	0	0	0	0
HRVOrtho5minLnTP	0	0	0	0	0.54
HRVOrtho5minLnVLF	0	0	0	0	0
HRVOrtho5minLnLF	0	0	0	0	0
HRVOrtho5minLnHF	0	0	0	0	0.29
QST.SHRVSTROOPColor5MinMeanHR	0	0	0.48	0	0
QST.SHRVSTROOPColor5MinSDNN	0	0.76	0	0	0
QST.SHRVSTROOPColor5MinRMSSD	0	0.42	0	0	0
HRVStroopColor5minLnTP	0	0	0	0	0
HRVStroopColor5minLnVLF	0	0	0	0	0
HRVStroopColor5minLnLF	0	0	0	0	0
HRVStroopColor5minLnHF	0	0	0	0	0
QST.SHRVSTROOPEmotion5mnMeanHR	0	0	0.58	0	0
QST.SHRVSTROOPEmotion5mnSDNN	0	0.42	0	0	0
QST.SHRVSTROOPEmotion5mnRMSSD	0	0.27	0	0	0
HRVStroopEmotion5minLnTP	0	0	0	0	0
HRVStroopEmotion5minLnVLF	0	0	0	0	0
Percent Variance Explained	0.14	0.21	0.3	0.34	0.4

Table A2.14: PCA/SFA QST Results for Cases, c=0.4

Component	1	2	3	4	5
Pressure Pain Threshhold					
Temporalis	0	0	0	0.47	0
Masseter	0	0	0	0.5	0
TMJ	0	0	0	0.5	0
Trapezius	0	0	0	0.4	0
Epicondyl	0	0	0	0.36	0
Mechanical Cutaneous Pain Threshold	0	0	0	0	0
Single Stimulus					
256 mN Probe	0	0	0	0	0.54
512 mN Probe	0	0	0	0	0.54
Temporal Summation					
256 mN probe	0	0	0	0	0
512 mN probe	0	0.07	0	0	0
After Sensation Ratings					
15 s, 256 mN probe	0.12	0	0	0	0.4
15 s, 512 mN probe	0.15	0	0	0	0.32
30 s, 256 mN probe	0.16	0	0	0	0.29
30 s, 512 mN probe	0.21	0	0	0	0.27
Heat Pain Tolerance	0	-0.05	0	0	0
Single Stimulus					
46	0	0.28	-0.06	0	0
48	0	0.36	-0.09	0	0
50	0	0.42	0	0	0
Area Under the Curve					
46	0	0.42	0	0	0
48	0	0.48	0	0	0
50	0	0.44	0	0	0
After Sensations					
15 s, 46	0.37	0	0	0	0
15 s, 48	0.38	0	0	0	0
15 s, 50	0.34	0	0	0	0
30 s, 46	0.39	0	0	0	0
30 s, 48	0.42	0	0	0	0
30 s, 50	0.41	0	0	0	0
Temporal Summation: Highest Minus First Rating					
46	0	0	0.42	0	0
48	0	0	0.48	0	0
50	0	-0.02	0.32	0	0
Temporal Summation: Slope of Line for First 3 Ratings					
46	0	0	0.32	0	0
48	0	0	0.45	0	0
50	0	0	0.41	0	0
Percent Variance Explained	0.2	0.35	0.48	0.6	0.69

Table A2.15: Autonomic Results for the OPPERA Case Control Study, c=0.4

Component	1	2	3	4	5
AvgRestingSPB	0.29	0	0	0	0
AvgRestingDPB	0.22	0	0	0	0
AvgRestingMAP	0.29	0	0	0	0
AvgRestingHR	0	0	0.3	0	0
InitialOrthoSPB	0.18	0	0	0	0
InitialOrthoDPB	0.21	0	0	0	0
InitialOrthoMAP	0.26	0	0	0	0
InitialOrthoHR	0	0	0.21	0	0
StroopColorSPBmean	0.33	0	0	0	0
StroopColorDPBmean	0.31	0	0	0	0
StroopColorMAPmean	0.35	0	0	0	0
StroopColorHRmean	0	0	0.4	0	0
StroopEmotionalSPBmean	0.33	0	0	0	0
StroopEmotionalDPBmean	0.29	0	0	0	0
StroopEmotionalMAPmean	0.35	0	0	0	0
StroopEmotionalHRmean	0	0	0.41	0	0
QST.SHRVBaseline20MinMeanHR	0	0	0.3	0	0
QST.SHRVBaseline20MinSDNN	0	0	0	0.38	0
QST.SHRVBaseline20MinRMSSD	0	0	0	0.18	0
HRVBaseline20minLnTP	0	0	0	0.52	0
HRVBaseline20minLnVLF	0	0	0	0.38	0
HRVBaseline20minLnLF	0	0	0	0.48	0
HRVBaseline20minLnHF	0	0	0	0.42	0
QST.SHRVOrthostatic5MinMeanHR	0	0	0.34	0	0
QST.SHRVOrthostatic5MinSDNN	0	0	0	0	0.47
QST.SHRVOrthostatic5MinRMSSD	0	0	0	0	0.41
HRVOrtho5minLnTP	0	0	0	0	0.43
HRVOrtho5minLnVLF	0	0	0	0	0.31
HRVOrtho5minLnLF	0	0	0	0	0.37
HRVOrtho5minLnHF	0	0	0	0	0.44
QST.SHRVSTROOPColor5MinMeanHR	0	0	0.41	0	0
QST.SHRVSTROOPColor5MinSDNN	0	0.42	0	0	0
QST.SHRVSTROOPColor5MinRMSSD	0	0.4	0	0	0
HRVStroopColor5minLnTP	0	0.34	0	0	0
HRVStroopColor5minLnVLF	0	0.2	0	0	0
HRVStroopColor5minLnLF	0	0.3	0	0	0
HRVStroopColor5minLnHF	0	0.33	0	0	0
QST.SHRVSTROOPEmotion5mnMeanHR	0	0	0.42	0	0
QST.SHRVSTROOPEmotion5mnSDNN	0	0.35	0	0	0
QST.SHRVSTROOPEmotion5mnRMSSD	0	0.32	0	0	0
HRVStroopEmotion5minLnTP	0	0.25	0	0	0
HRVStroopEmotion5minLnVLF	0	0.12	0	0	0
Percent Variance Explained	0.2	0.36	0.51	0.61	0.72

Table A2.16: Autonomic Results for the OPPERA Case Control Study, c=0.5

Component	1	2	3	4	5
AvgRestingSPB	0.28	0	0	0	0
AvgRestingDPB	0.2	0	0	0	0
AvgRestingMAP	0.29	0	0	0	0
AvgRestingHR	0	0	0.28	0	0
InitialOrthoSPB	0.14	0	0	0	0
InitialOrthoDPB	0.18	0	0	0	0
InitialOrthoMAP	0.25	0	0	0	0
InitialOrthoHR	0	0	0.17	0	0
StroopColorSPBmean	0.34	0	0	0	0
StroopColorDPBmean	0.31	0	0	0	0
StroopColorMAPmean	0.37	0	0	0	0
StroopColorHRmean	0	0	0.4	0	0
StroopEmotionalSPBmean	0.34	0	0	0	0
StroopEmotionalDPBmean	0.29	0	0	0	0
StroopEmotionalMAPmean	0.36	0	0	0	0
StroopEmotionalHRmean	0	0	0.42	0	0
QST.SHRVBaseline20MinMeanHR	0	0	0.28	0	0
QST.SHRVBaseline20MinSDNN	0	0	0	0.37	0
QST.SHRVBaseline20MinRMSSD	0	0	0	0.11	0
HRVBaseline20minLnTP	0	0	0	0.55	0
HRVBaseline20minLnVLF	0	0	0	0.37	0
HRVBaseline20minLnLF	0	0	0	0.49	0
HRVBaseline20minLnHF	0	0	0	0.42	0
QST.SHRVOrthostatic5MinMeanHR	0	0	0.33	0	0
QST.SHRVOrthostatic5MinSDNN	0	0	0	0	0.49
QST.SHRVOrthostatic5MinRMSSD	0	0	0	0	0.41
HRVOrtho5minLnTP	0	0	0	0	0.44
HRVOrtho5minLnVLF	0	0	0	0	0.28
HRVOrtho5minLnLF	0	0	0	0	0.36
HRVOrtho5minLnHF	0	0	0	0	0.44
QST.SHRVSTROOPColor5MinMeanHR	0	0	0.42	0	0
QST.SHRVSTROOPColor5MinSDNN	0	0.45	0	0	0
QST.SHRVSTROOPColor5MinRMSSD	0	0.43	0	0	0
HRVStroopColor5minLnTP	0	0.34	0	0	0
HRVStroopColor5minLnVLF	0	0.16	0	0	0
HRVStroopColor5minLnLF	0	0.29	0	0	0
HRVStroopColor5minLnHF	0	0.33	0	0	0
QST.SHRVSTROOPEmotion5mnMeanHR	0	0	0.44	0	0
QST.SHRVSTROOPEmotion5mnSDNN	0	0.36	0	0	0
QST.SHRVSTROOPEmotion5mnRMSSD	0	0.32	0	0	0
HRVStroopEmotion5minLnTP	0	0.22	0	0	0
HRVStroopEmotion5minLnVLF	0	0.05	0	0	0
Percent Variance Explained	0.2	0.35	0.5	0.6	0.71

Table A2.17: Autonomic Results for the OPPERA Case Control Study, c=0.6

Component	1	2	3	4	5
AvgRestingSPB	0.28	0	0	0	0
AvgRestingDPB	0.16	0	0	0	0
AvgRestingMAP	0.29	0	0	0	0
AvgRestingHR	0	0	0.25	0	0
InitialOrthoSPB	0.08	0	0	0	0
InitialOrthoDPB	0.14	0	0	0	0
InitialOrthoMAP	0.22	0	0	0	0
InitialOrthoHR	0	0	0.08	0	0
StroopColorSPBmean	0.36	0	0	0	0
StroopColorDPBmean	0.31	0	0	0	0
StroopColorMAPmean	0.4	0	0	0	0
StroopColorHRmean	0	0	0.42	0	0
StroopEmotionalSPBmean	0.35	0	0	0	0
StroopEmotionalDPBmean	0.29	0	0	0	0
StroopEmotionalMAPmean	0.39	0	0	0	0
StroopEmotionalHRmean	0	0	0.44	0	0
QST.SHRVBaseline20MinMeanHR	0	0	0.24	0	0
QST.SHRVBaseline20MinSDNN	0	0	0	0.33	0
QST.SHRVBaseline20MinRMSSD	0	0	0	0	0
HRVBaseline20minLnTP	0	0	0	0.59	0
HRVBaseline20minLnVLF	0	0	0	0.34	0
HRVBaseline20minLnLF	0	0	0	0.51	0
HRVBaseline20minLnHF	0	0	0	0.41	0
QST.SHRVOrthostatic5MinMeanHR	0	0	0.31	0	0
QST.SHRVOrthostatic5MinSDNN	0	0	0	0	0.52
QST.SHRVOrthostatic5MinRMSSD	0	0	0	0	0.41
HRVOrtho5minLnTP	0	0	0	0	0.45
HRVOrtho5minLnVLF	0	0	0	0	0.21
HRVOrtho5minLnLF	0	0	0	0	0.33
HRVOrtho5minLnHF	0	0	0	0	0.45
QST.SHRVSTROOPColor5MinMeanHR	0	0	0.44	0	0
QST.SHRVSTROOPColor5MinSDNN	0	0.49	0	0	0
QST.SHRVSTROOPColor5MinRMSSD	0	0.46	0	0	0
HRVStroopColor5minLnTP	0	0.34	0	0	0
HRVStroopColor5minLnVLF	0	0.07	0	0	0
HRVStroopColor5minLnLF	0	0.26	0	0	0
HRVStroopColor5minLnHF	0	0.32	0	0	0
QST.SHRVSTROOPEmotion5mnMeanHR	0	0	0.46	0	0
QST.SHRVSTROOPEmotion5mnSDNN	0	0.36	0	0	0
QST.SHRVSTROOPEmotion5mnRMSSD	0	0.31	0	0	0
HRVStroopEmotion5minLnTP	0	0.16	0	0	0
HRVStroopEmotion5minLnVLF	0	0	0	0	0
Percent Variance Explained	0.19	0.33	0.47	0.56	0.67

Table A2.18: Autonomic Results for the OPPERA Case Control Study, c=0.8

Component	1	2	3	4	5
AvgRestingSPB	0.14	0	0	0	0
AvgRestingDPB	0	0	0	0	0
AvgRestingMAP	0.17	0	0	0	0
AvgRestingHR	0	0	0	0	0
InitialOrthoSPB	0	0	0	0	0
InitialOrthoDPB	0	0	0	0	0
InitialOrthoMAP	0	0	0	0	0
InitialOrthoHR	0	0	0	0	0
StroopColorSPBmean	0.4	0	0	0	0
StroopColorDPBmean	0.26	0	0	0	0
StroopColorMAPmean	0.54	0	0	0	0
StroopColorHRmean	0	0	0.42	0	0
StroopEmotionalSPBmean	0.38	0	0	0	0
StroopEmotionalDPBmean	0.18	0	0	0	0
StroopEmotionalMAPmean	0.51	0	0	0	0
StroopEmotionalHRmean	0	0	0.51	0	0
QST.SHRVBaseline20MinMeanHR	0	0	0	0	0
QST.SHRVBaseline20MinSDNN	0	0	0	0	0
QST.SHRVBaseline20MinRMSSD	0	0	0	0	0
HRVBaseline20minLnTP	0	0	0	0.84	0
HRVBaseline20minLnVLF	0	0	0	0	0
HRVBaseline20minLnLF	0	0	0	0.53	0
HRVBaseline20minLnHF	0	0	0	0.08	0
QST.SHRVOrthostatic5MinMeanHR	0	0	0.03	0	0
QST.SHRVOrthostatic5MinSDNN	0	0	0	0	0.8
QST.SHRVOrthostatic5MinRMSSD	0	0	0	0	0.17
HRVOrtho5minLnTP	0	0	0	0	0.39
HRVOrtho5minLnVLF	0	0	0	0	0
HRVOrtho5minLnLF	0	0	0	0	0
HRVOrtho5minLnHF	0	0	0	0	0.42
QST.SHRVSTROOPColor5MinMeanHR	0	0	0.49	0	0
QST.SHRVSTROOPColor5MinSDNN	0	0.77	0	0	0
QST.SHRVSTROOPColor5MinRMSSD	0	0.62	0	0	0
HRVStroopColor5minLnTP	0	0.03	0	0	0
HRVStroopColor5minLnVLF	0	0	0	0	0
HRVStroopColor5minLnLF	0	0	0	0	0
HRVStroopColor5minLnHF	0	0	0	0	0
QST.SHRVSTROOPEmotion5mnMeanHR	0	0	0.57	0	0
QST.SHRVSTROOPEmotion5mnSDNN	0	0.13	0	0	0
QST.SHRVSTROOPEmotion5mnRMSSD	0	0	0	0	0
HRVStroopEmotion5minLnTP	0	0	0	0	0
HRVStroopEmotion5minLnVLF	0	0	0	0	0
Percent Variance Explained	0.14	0.19	0.28	0.33	0.39

Table A2.19: PCA/SFA Psychosocial Results for the Entire OPPERA Cohort, c=0.4

Component	1	2	3	4
KOHN.Global.Score	0	0	0.12	0
PILL.Global.Score	0.28	0	0	0
POMS.PositiveAffectScore	0	-0.52	0	0
POMS.NegativeAffectScore	0.11	0.23	0	0
PSS.PerceivedStress	0	0.39	0	0
SCL.90R.Depression	0.48	0	0	0
SCL.90R.Somatization	0.47	0	0	0
SCL.90R.Anxiety	0.5	0	0	0
SCL.90R.Hostility	0.45	0	0	0
CSQ.Distraction	0	0	0	0.48
CSQ.IgnoringPain	0	0	0	0.49
CSQ.Distancing	0	0	0	0.49
CSQ.Coping	0	0	0	0.54
CSQ.Praying	0	0	0.21	0
STAIY1.StateTraitAnxiety	0	0.48	0	0
STAIY2.StateTraitAnxiety	0.04	0.48	0	0
PCS.Rumination	0	0	0.61	0
PCS.Magnification	0	0	0.5	0
PCS.Helplessness	0	0	0.57	0
EPQ.Escale	0	-0.08	0	0
EPQ.Nscale	0	0.24	0	0
Percent Variance Explained	0.17	0.34	0.46	0.57

Table A2.20: PCA/SFA Psychosocial Results for the Entire OPPERA Cohort, c=0.5

Component	1	2	3	4
KOHN.Global.Score	0	0	0	0
PILL.Global.Score	0.22	0	0	0
POMS.PositiveAffectScore	0	-0.56	0	0
POMS.NegativeAffectScore	0	0.14	0	0
PSS.PerceivedStress	0	0.36	0	0
SCL.90R.Depression	0.49	0	0	0
SCL.90R.Somatization	0.48	0	0	0
SCL.90R.Anxiety	0.52	0	0	0
SCL.90R.Hostility	0.45	0	0	0
CSQ.Distraction	0	0	0	0.46
CSQ.IgnoringPain	0	0	0	0.49
CSQ.Distancing	0	0	0	0.48
CSQ.Coping	0	0	0	0.56
CSQ.Praying	0	0	0.08	0
STAIY1.StateTraitAnxiety	0	0.5	0	0
STAIY2.StateTraitAnxiety	0	0.51	0	0
PCS.Rumination	0	0	0.65	0
PCS.Magnification	0	0	0.49	0
PCS.Helplessness	0	0	0.58	0
EPQ.Escale	0	0	0	0
EPQ.Nscale	0	0.15	0	0
Percent Variance Explained	0.16	0.31	0.43	0.54

Table A2.21: PCA/SFA Psychosocial Results for the Entire OPPERA Cohort, c=0.7

Component	1	2	3	4
KOHN.Global.Score	0	0	0	0
PILL.Global.Score	0	0	0	0
POMS.PositiveAffectScore	0	-0.76	0	0
POMS.NegativeAffectScore	0	0	0	0
PSS.PerceivedStress	0	0	0	0
SCL.90R.Depression	0.5	0	0	0
SCL.90R.Somatization	0.48	0	0	0
SCL.90R.Anxiety	0.61	0	0	0
SCL.90R.Hostility	0.39	0	0	0
CSQ.Distraction	0	0	0	0.21
CSQ.IgnoringPain	0	0	0	0.38
CSQ.Distancing	0	0	0	0.35
CSQ.Coping	0	0	0	0.83
CSQ.Praying	0	0	0	0
STAIY1.StateTraitAnxiety	0	0.43	0	0
STAIY2.StateTraitAnxiety	0	0.48	0	0
PCS.Rumination	0	0	0.78	0
PCS.Magnification	0	0	0.25	0
PCS.Helplessness	0	0	0.57	0
EPQ.Escale	0	0	0	0
EPQ.Nscale	0	0	0	0
Percent Variance Explained	0.14	0.24	0.35	0.44

Table A2.22: PCA/SFA Psychosocial Results for the Entire OPPERA Cohort, c=0.4 to 0.55

Component	1	2	3	4
KOHN.Global.Score	0	0	0	0
PILL.Global.Score	0.28	0	0	0
POMS.PositiveAffectScore	0	-0.54	0	0
POMS.NegativeAffectScore	0.11	0.2	0	0
PSS.PerceivedStress	0	0.38	0	0
SCL.90R.Depression	0.48	0	0	0
SCL.90R.Somatization	0.47	0	0	0
SCL.90R.Anxiety	0.5	0	0	0
SCL.90R.Hostility	0.45	0	0	0
CSQ.Distraction	0	0	0	0.45
CSQ.IgnoringPain	0	0	0	0.49
CSQ.Distancing	0	0	0	0.48
CSQ.Coping	0	0	0	0.57
CSQ.Praying	0	0	0.08	0
STAIY1.StateTraitAnxiety	0	0.49	0	0
STAIY2.StateTraitAnxiety	0.04	0.5	0	0
PCS.Rumination	0	0	0.65	0
PCS.Magnification	0	0	0.49	0
PCS.Helplessness	0	0	0.58	0
EPQ.Escale	0	-0.01	0	0
EPQ.Nscale	0	0.2	0	0
Percent Variance Explained	0.17	0.33	0.45	0.55

Table A2.23: PCA/SFA Psychosocial Results for the Entire OPPERA Cohort, c=0.6 to 0.45

Component	1	2	3	4
KOHN.Global.Score	0	0	0	0
PILL.Global.Score	0.09	0	0	0
POMS.PositiveAffectScore	0	-0.59	0	0
POMS.NegativeAffectScore	0	0.04	0	0
PSS.PerceivedStress	0	0.33	0	0
SCL.90R.Depression	0.5	0	0	0
SCL.90R.Somatization	0.49	0	0	0
SCL.90R.Anxiety	0.55	0	0	0
SCL.90R.Hostility	0.45	0	0	0
CSQ.Distraction	0	0	0	0.47
CSQ.IgnoringPain	0	0	0	0.49
CSQ.Distancing	0	0	0	0.49
CSQ.Coping	0	0	0	0.55
CSQ.Praying	0	0	0.08	0
STAIY1.StateTraitAnxiety	0	0.51	0	0
STAIY2.StateTraitAnxiety	0	0.52	0	0
PCS.Rumination	0	0	0.65	0
PCS.Magnification	0	0	0.49	0
PCS.Helplessness	0	0	0.58	0
EPQ.Escale	0	0	0	0
EPQ.Nscale	0	0.05	0	0
Percent Variance Explained	0.15	0.29	0.41	0.51

Table A2.24: PCA/SFA Psychosocial Results for the Entire OPPERA Cohort, c=0.7 to 0.4

Component	1	2	3	4
KOHN.Global.Score	0	0	0	0
PILL.Global.Score	0	0	0	0
POMS.PositiveAffectScore	0	-0.62	0	0
POMS.NegativeAffectScore	0	0	0	0
PSS.PerceivedStress	0	0.27	0	0
SCL.90R.Depression	0.5	0	0	0
SCL.90R.Somatization	0.48	0	0	0
SCL.90R.Anxiety	0.61	0	0	0
SCL.90R.Hostility	0.39	0	0	0
CSQ.Distraction	0	0	0	0.48
CSQ.IgnoringPain	0	0	0	0.49
CSQ.Distancing	0	0	0	0.49
CSQ.Coping	0	0	0	0.54
CSQ.Praying	0	0	0.08	0
STAIY1.StateTraitAnxiety	0	0.51	0	0
STAIY2.StateTraitAnxiety	0	0.53	0	0
PCS.Rumination	0	0	0.65	0
PCS.Magnification	0	0	0.49	0
PCS.Helplessness	0	0	0.58	0
EPQ.Escale	0	0	0	0
EPQ.Nscale	0	0	0	0
Percent Variance Explained	0.14	0.27	0.39	0.49

BIBLIOGRAPHY

[1] Bair, E., Brownstein, N. C., Ohrbach, R., Greenspan, J. D., Dubner, R., Fillingim, R. B., Maixner, W., Smith, S., Diatchenko, L., Gonzalez, Y., Gordon, S., Lim, P.-F., Ribeiro-Dasilva, M., Dampier, D., Knott, C., and Slade, G. D. (2013). Study protocol, sample characteristics and loss-to-follow-up: the OPPERA prospective cohort study. In press.

[2] Baker, S. G. (1994). Regression analysis of grouped survival data with incomplete covariates: Nonignorable missing-data and censoring mechanisms. *Biometrics*, 50(3):pp. 821–826.

[3] Bartholomew, D. J. (1984). The foundations of factor analysis. *Biometrika*, 71(2):pp. 221–232.

[4] Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24:1713–1723.

[5] Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):pp. 1196–1217.

[6] Chen, P., He, R., Shen, J.-s., and Sun, J.-g. (2009). Regression analysis of right-censored failure time data with missing censoring indicators. *Acta Mathematicae Applicatae Sinica (English Series)*, 25:415–426. 10.1007/s10255-008-8807-1.

[7] Cho, M. and Schenker, N. (1999). Fitting the log-f accelerated failure time model with incomplete covariate data. *Biometrics*, 55(3):pp. 826–833.

[8] Cook, T. D. and Kosorok, M. R. (2004). Analysis of time-to-event data with incomplete event adjudication. *Journal of the American Statistical Association*, 99(468):pp. 1140–1152.

[9] Cooper, J. C. B. (1983). Factor analysis: An overview. *The American Statistician*, 37(2):pp. 141–147.

[10] Cox, D. R. (1972). Models and life-tables. *Journal of the Royal Statistical Society. Series (B) (Methodological)*, 34(2):187–220.

[11] Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

[12] Craiu, R. V. and Duchesne, T. (2004). Inference based on the em algorithm for the competing risks model with masked causes of failure. *Biometrika*, 91(3):pp. 543–558.

[13] Dafni, U. G. and Tsiatis, A. A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54(4):1445–1462.

[14] Davison, A. C., Hinkley, D. V., and Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science*, 18(2):pp. 141–157.

[15] De Gruttola, V. and Tu, X. M. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, 50(4):1003–1014.

[16] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):pp. 1–38.

[17] Dewanji, A. (1992). A note on a test for competing risks with missing failure type. *Biometrika*, 79(4):pp. 855–857.

[18] Dewanji, A. and Sengupta, D. (2003). Estimation of competing risks with general missing pattern in failure types. *Biometrics*, 59(4):pp. 1063–1070.

[19] DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3):pp. 189–212.

[20] Dinse, G. E. (1982). Nonparametric estimation for partially-complete time and type of failure data. *Biometrics*, 38(2):pp. 417–431.

[21] Dinse, G. E. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *Journal of the American Statistical Association*, 81(394):pp. 328–336.

[22] Dupuy, J.-F., Grama, I., and Mesbah, M. (2006). Asymptotic theory for the cox model with missing time-dependent covariate. *The Annals of Statistics*, 34(2):903–924.

[23] Dupuy, J.-F. and Mesbah, M. (2002). Joint modeling of event time and nonignorable missing longitudinal data. *Lifetime Data Analysis*, 8:99–115.

[24] Dupuy, J.-F. and Mesbah, M. (2004). Estimation of the asymptotic variance of semiparametric maximum likelihood estimators in the cox model with a missing time-dependent covariate. *Communications in Statistics - Theory and Methods*, 33:1385–1401.

[25] Dworkin, S. and LeResche, L. (1992). Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. *Journal of Craniomandibular Disorders*, 6(4):301–355.

[26] Efron, B. (1977). Efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565.

[27] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

[28] Efron, B. (1981a). Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374):pp. 312–319.

[29] Efron, B. (1981b). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):pp. 589–599.

[30] Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426):pp. 463–475.

[31] Efron, B. (2003). Second thoughts on the bootstrap. *Statistical Science*, 18(2):pp. 135–140.

[32] Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):pp. 54–75.

[33] Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman and Hall/CRC, Boca Raton, FL.

[34] Faucett, C. L., Schenker, N., and Taylor, J. M. G. (2002). Survival analysis using auxiliary variables via multiple imputation, with application to aids clinical trial data. *Biometrics*, 58(1):37–47.

[35] Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A gibbs sampling approach. *Statistics in Medicine*, 15:1663–1685.

[36] Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, 21(4):826–838.

[37] Fillingim, R. B., Ohrbach, R., Greenspan, J. D., Knott, C., Dubner, R., Bair, E., Baraian, C., Slade, G. D., and Maixner, W. (2011a). Potential psychosocial risk factors for chronic TMD: Descriptive data and empirically identified domains from the OPPERA case-control study. *The Journal of Pain*, 12(11, Supplement):T46 – T60.

[38] Fillingim, R. B., Slade, G. D., Diatchenko, L., Dubner, R., Greenspan, J. D., Knott, C., Ohrbach, R., and Maixner, W. (2011b). Summary of findings from the OPPERA baseline case-control study: Implications and future directions. *The Journal of Pain*, 12(11, Supplement):T102–T107.

[39] Flehinger, B. J., Reiser, B., and Yashchin, E. (1998). Survival with competing risks and masked causes of failures. *Biometrika*, 85(1):pp. 151–164.

[40] Fleming, T. R. and Lin, D. Y. (2000). Survival analysis in clinical trials: Past development and future directions. *Biometrics*, 56(4):971–973.

[41] Gao, G. and Tsiatis, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika*, 92(4):pp. 875–891.

[42] Gaynor, S. and Bair, E. (2013). Sheila's clustering algorithm. *Journal of Incomplete Papers*, 1(3):1–25.

[43] Gijbels, I., Lin, D., and Ying, Z. (2007). Non- and semi- parametric analysis of failure-time data with missing failure indicators. *Lecture Notes-Monograph Series*, 54(1):203–223.

[44] Goetghebeur, E. and Ryan, L. (1990). A modified log rank test for competing risks with missing failure type. *Biometrika*, 77(1):pp. 207–211.

[45] Goetghebeur, E. and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika*, 82(4):821–833.

[46] Greenspan, J. D., Slade, G. D., Bair, E., Dubner, R., Fillingim, R. B., Ohrbach, R., Knott, C., Mulkey, F., Rothwell, R., and Maixner, W. (2011). Pain sensitivity risk factors for chronic TMD: Descriptive data and empirically identified domains from the OPPERA case control study. *The Journal of Pain*, 12(11, Supplement):T61 – T74.

[47] He, J., Li, H., Edmondson, A. C., Rader, D. J., and Li, M. (2011). A gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics*.

[48] Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

[49] Herring, A. H. and Ibrahim, J. G. (2001). Likelihood-based methods for missing covariates in the cox proportional hazards model. *Journal of the American Statistical Association*, 96(453):292–302.

[50] Hoel, D. G. (1972). A representation of mortality data by competing risks. *Biometrics*, 28(2):pp. 475–488.

[51] Horton, N. J., Lipsitz, S. R., and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, 57(4):pp. 229–232.

[52] Hu, P., Tsiatis, A. A., and Davidian, M. (1998). Estimating the parameters in the cox model when covariate variables are measured with error. *Biometrics*, 54(4):pp. 1407–1419.

[53] Huang, C.-Y. and Wang, M.-C. (2004). Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association*, 99(468):pp. 1153–1165.

[54] Huang, X. and Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics*, 58(3):pp. 510–520.

[55] Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., and Williams, L. J. (1997). Exploratory and confirmatory factor analysis: guidelines, issues, and alternatives. *Journal of Organizational Behavior*, 18(6):667–683.

[56] Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica*, 14:863–883.

[57] Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28:2796–2801.

[58] Janal, M. N., Raphael, K. G., Nayak, S., and Klausner, J. (2008). Prevalence of myofascial temporomandibular disorder in us community women. *Journal of Oral Rehabilitation*, 35(11):801–809.

[59] Kalbfleish and Prentice (1979). Hazard rate models with covariates. *Biometrics*, 35(1):25–39.

[60] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

[61] Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, 35(1):pp. 139–156.

[62] Lee, S. M. S. and Young, G. A. (2003). Prepivoting by weighted bootstrap iteration. *Biometrika*, 90(2):pp. 393–410.

[63] Li, H., Gail, M. H.and Berndt, S., and Chatterjee, N. (2010). Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genetic Epidemiology*, 34:427–433.

[64] Lin, D. Y., Robins, J. M., and Wei, L. J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika*, 83(2):pp. 381–393.

[65] Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, 88(424):pp. 1341–1349.

[66] Lin, D. Y. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*, 33:256–265.

[67] Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97:53–65.

[68] Link, C. L. (1984). Confidence intervals for the survival function using cox's proportional hazard model with covariates. *Biometrics*, 40(3):601–609.

[69] Link, W. A. (1989). A model for informative censoring. *Journal of the American Statistical Association*, 84(407):pp. 749–752.

[70] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Mathematical Statistics.

[71] Lo, S.-H. (1991). Estimating a survival function with incomplete cause-of-death data. *Journal of Multivariate Analysis*, 39(2):217–235.

[72] Louis, T. A. (1981). Nonparametric analysis of an accelerated failure time model. *Biometrika*, 68(2):pp. 381–390.

[73] Lu, K. and Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*, 57(4):pp. 1191–1197.

[74] Magder, L. S. and Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, 146(2):195–203.

[75] Maixner, W., Diatchenko, L., Dubner, R., Fillingim, R. B., Greenspan, J. D., Knott, C., Ohrbach, R., Weir, B., and Slade, G. D. (2011a). Orofacial pain prospective evaluation and risk assessment study - the OPPERA study. *The Journal of Pain*, 12(11):T4–T11.

[76] Maixner, W., Greenspan, J. D., Dubner, R., Bair, E., Mulkey, F., Miller, V., Knott, C., Slade, G. D., Ohrbach, R., Diatchenko, L., and Fillingim, R. B. (2011b). Potential autonomic risk factors for chronic TMD: Descriptive data and empirically identified domains from the OPPERA case-control study. *The Journal of Pain*, 12(11, Supplement):T75 – T91.

[77] Matsui, S. (2004). Analysis of times to repeated events in two-arm randomized trials with noncompliance and dependent censoring. *Biometrics*, 60(4):pp. 965–976.

[78] McKeague, I. W. and Subramanian, S. (1998). Product-limit estimators and Cox regression with missing censoring information. *Scandinavian Journal of Statistics*, 25(4):pp. 589–601.

[79] Meier, A. S., Richardson, B. A., and Hughes, J. P. (2003). Discrete proportional hazards models for mismeasured outcomes. *Biometrics*, 59(4):pp. 947–954.

[80] Miloslavsky, M., Keles, S., Laan, M. J. v. d., and Butler, S. (2004). Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(1):pp. 239–257.

[81] Moeschberger, M. L. and David, H. A. (1971). Life tests under competing causes of failure and the theory of competing risks. *Biometrics*, 27(4):pp. 909–933.

[82] Monsees, G. M., Tamimi, R. M., and Kraft, P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology*, 33:717–728.

[83] Mori, M., Woolson, R. F., and Woodworth, G. G. (1994). Slope estimation in the presence of informative right censoring: Modeling the number of observations as a geometric random variable. *Biometrics*, 50(1):pp. 39–50.

[84] Mukerjee, H. and Wang, J.-L. (1993). Nonparametric maximum likelihood estimation of an increasing hazard rate for uncertain cause-of-death data. *Scandinavian Journal of Statistics*, 20(1):pp. 17–33.

[85] Nair, V. N. (1993). Bounds for reliability estimation under dependent censoring. *International Statistical Review / Revue Internationale de Statistique*, 61(1):pp. 169–182.

[86] Ohrbach, R., Fillingim, R. B., Mulkey, F., Gonzalez, Y., Gordon, S., Gremillion, H., Lim, P.-F., Ribeiro-Dasilva, M., Greenspan, J. D., Knott, C., Maixner, W., and Slade, G. (2011a). Clinical findings and pain symptoms as potential risk factors for chronic TMD: Descriptive data and empirically identified domains from the OPPERA case-control study. *The Journal of Pain*, 12(11, Supplement):T27 – T45.

[87] Ohrbach, R., Fillingim, R. B., Mulkey, F., Gonzalez, Y., Gordon, S., Gremillion, H., Lim, P.-F., Ribeiro-Dasilva, M., Greenspan, J. D., Knott, C., Maixner, W., and Slade, G. (2011b). Clinical findings and pain symptoms as potential risk factors for chronic TMD: Descriptive data and empirically identified domains from the OPPERA case-control study. *The Journal of Pain*, 12(11, Supplement):T27 – T45.

[88] O'Sullivan, F. (1993). Nonparametric estimation in the cox model. *The Annals of Statistics*, 21(1):pp. 124–145.

[89] Paik, M. C. and Tsai, W.-Y. (1997). On using the cox proportional hazards model with missing covariates. *Biometrika*, 84(3):579–593.

[90] Paulino, C. D. M. and Pereira, C. A. d. B. (1995). Bayesian methods for categorical data under informative general censoring. *Biometrika*, 82(2):pp. 439–446.

[91] Pohlmann, J. T. (2004). Use and interpretation of factor analysis in "the journal of educational research": 19922002. *The Journal of Educational Research*, 98(1):pp. 14–22.

[92] Potscher, B. M. and Prucha, I. R. (1989). A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica*, 57(3):pp. 675–683.

[93] Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):pp. 331–342.

[94] Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., J., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):pp. 541–554.

[95] Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):pp. 403–411.

[96] Racine-Poon, A. H. and Hoel, D. G. (1984). Nonparametric estimation of the survival function when cause of death is uncertain. *Biometrics*, 40(4):pp. 1151–1158.

[97] Richardson, D. B., Rzehak, P., Klenk, J., and Weiland, S. K. (2007). Analyses of case-control data for additional outcomes. *Epidemiology*, 18:441–445.

[98] Robins, J. M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharaceutical Section*, pages 24–33. American Statistical Association.

[99] Robins, J. M. and Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779–788.

[100] Rotnitzky, A., Scharfstein, D., Su, T.-L., and Robins, J. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics*, 57(1):pp. 103–113.

[101] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):pp. 581–592.

[102] Rubin, D. B. (1981). The bayesian bootstrap. *The Annals of Statistics*, 9(1):pp. 130–134.

[103] Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):pp. 473–489.

[104] Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15(11):1576–1583.

[105] Scharfstein, D., Robins, J. M., Eddings, W., and Rotnitzky, A. (2001). Inference in randomized studies with informative censoring and discrete time-to-event endpoints. *Biometrics*, 57(2):pp. 404–413.

[106] Scharfstein, D. O. and Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):pp. 617–634.

[107] Schaubel, D. E. and Cai, J. (2006). Multiple imputation methods for recurrent event data with missing event category. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, (4):677–692.

[108] Sen, P. K. (1995). Censoring in theory and practice: Statistical perspectives and controversies. *Lecture Notes-Monograph Series*, 27:pp. 177–192.

[109] Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015 – 1034.

[110] Siannis, F. (2004). Applications of a parametric model for informative censoring. *Biometrics*, 60(3):pp. 704–714.

[111] Singh, K. (1981). On the asymptotic accuracy of efron's bootstrap. *The Annals of Statistics*, 9(6):pp. 1187–1195.

[112] Singh, K. and Babu, G. J. (1990). On asymptotic optimality of the bootstrap. *Scandinavian Journal of Statistics*, 17(1):pp. 1–9.

[113] Slade, G. D., Bair, E., By, K., Mulkey, F., Baraian, C., Rothwell, R., Reynolds, M., Miller, V., Gonzalez, Y., Gordon, S., Ribeiro-Dasilva, M., Lim, P. F., Greenspan, J. D., Dubner, R., Fillingim, R. B., Diatchenko, L., Maixner, W., Dampier, D., Knott, C., and Ohrbach, R. (2011). Study methods, recruitment, sociodemographic findings, and demographic representativeness in the OPPERA study. *The Journal of Pain*, 12(11):T12–T26.

[114] Slade, G. D., Sanders, A., Bair, E., Brownstein, N. C., Fillingim, R. B., Maixner, W., Greenspan, J. D., and Ohrbach, R. (2013). Pre-clinical episodes of orofacial pain symptoms and their association with healthcare behaviors in the OPPERA prospective cohort study. *Pain*, 154:750–760.

[115] Smith, S. B., Maixner, D. W., Greenspan, J. D., Dubner, R., Fillingim, R. B., Ohrbach, R., Knott, C., Slade, G. D., Bair, E., Gibson, D. G., Zaykin, D. V., Weir, B. S., Maixner, W., and Diatchenko, L. (2011). Potential genetic risk factors for chronic TMD: Genetic associations from the OPPERA case control study. *The Journal of Pain*, 12(11, Supplement):T92 – T101.

[116] Snapinn, S. M. (1998). Survival analysis with uncertain endpoints. *Biometrics*, 54(1):pp. 209–218.

[117] Song, X., Davidian, M., and Tsiatis, A. A. (2002a). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics*, 3:511–528.

[118] Song, X., Davidian, M., and Tsiatis, A. A. (2002b). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58:742–753.

[119] Song, X., Sun, L., Mu, X., and Dinse, G. E. (2010). Additive hazards regression with censoring indicators missing at random. *Canadian Journal of Statistics*, 38(3):333–351.

[120] Song, X. and Wang, C. Y. (2008). Semiparametric approaches for joint modeling of longitudinal and survival data with time-varying coefficients. *Biometrics*, 64:557–566.

[121] Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified propotional hazard models. *Biometrika*, 73(2):363–369.

[122] Subramanian, S. (2000). Efficient estimation of regression coefficients and baseline hazard under proportionality of conditional hazards. *Journal of Statistical Planning and Inference*, 84(1-2):81 – 94.

[123] Subramanian, S. (2004). Asymptotically efficient estimation of a survival function in the missing censoring indicator model. *Journal of Nonparametric Statistics*, 16(5):797–817.

[124] Subramanian, S. (2006). Survival analysis for the missing censoring indicator model using kernel density estimation techniques. *Statistical Methodology*, 3(2):125 – 136.

[125] Subramanian, S. (2009). The multiple imputations based kaplan-meier estimator. *Statistics & Probability Letters*, 79(18):1906 – 1914.

[126] Subramanian, S. (2011). Multiple imputations and the missing censoring indicator model. *Journal of Multivariate Analysis*, 102(1):105 – 117.

[127] Subramanian, S. and Bandyopadhyay, D. (2010). Doubly robust semiparametric estimation for the missing censoring indicator model. *Statistics & Probability Letters*, 80(7-8):621 – 630.

[128] Subramanian, S. and Bean, D. (2008a). Hazard function estimation from homogeneous right censored data with missing censoring indicators. *Statistical Methodology*, 5(6):515 – 527.

[129] Subramanian, S. and Bean, D. (2008b). The missing censoring indicator model and the smoothed bootstrap. *Computational Statistics & Data Analysis*, 53(2):471 – 476.

[130] Tolonen, H., Helakorpi, S., Talala, K., Helasoja, V., Martelin, T., and Prttl, R. (2006). 25-year trends and socio-demographic differences in response rates: Finnish adult health behaviour survey. *European Journal of Epidemiology*, 21(6):pp. 409–415.

[131] Tsiatis, A. A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88:447–458.

[132] Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14:809–834.

[133] Tsiatis, A. A., Davidian, M., and McMeney, B. (2002a). Multiple imputation methods for testing treatment differences in survival distributions with missing cause of failure. *Biometrika*, 89(1):238–244.

[134] Tsiatis, A. A., Davidian, M., and McMeney, B. (2002b). Multiple imputation methods for testing treatment differences in survival distributions with missing cause of failure. *Biometrika*, 89(1):238–244.

[135] Tsiatis, A. A., De Gruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429):27–37.

[136] van der Laan, M. J. and McKeague, I. W. (1998). Efficient estimation from right-censored data when failure indicators are missing at random. *The Annals of Statistics*, 26(1):164–182.

[137] Velicer, W. F. and Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, 25(1):1–28.

[138] Wang, C.-Y. (2006). Corrected score estimator for joint modeling of longitudinal and failure time data. *Statistica Sinica*, 16:235–253.

[139] Wang, C. Y. and Chen, H. Y. (2001). Augmented inverse probability weighted estimator for cox missing covariate regression. *Biometrics*, 57(2):pp. 414–419.

[140] Wang, J. and Shete, S. (2011a). Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genetic Epidemiology*, 35:190–200.

[141] Wang, J. and Shete, S. (2011b). Power and type i error results for a bias-correction approach recently shown to provide accurate odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genetic Epidemiology*, 35:739–743.

[142] Wang, M.-C., Qin, J., and Chiang, C.-T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96(455):pp. 1057–1065.

[143] Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85(4):pp. 935–948.

[144] Wang, Q. and Dinse, G. (2011). Linear regression analysis of survival data with missing censoring indicators. *Lifetime Data Analysis*, 17:256–279. 10.1007/s10985-010-9175-8.

[145] Wang, Q., Liu, W., and Liu, C. (2009). Probability density estimation for survival data with censoring indicators missing at random. *Journal of Multivariate Analysis*, 100(5):835 – 850.

[146] Warren, L. L., Li, L., Nelson, M. R.and Ehm, M. G., Shen, J., Fraser, D., Aponte, J. L., Nangle, K. L., Slater, A. J., Woollard, P. M., Hall, M. D., Topp, S. D., Yuan, X., Cardon, L. R., Chissoe, S. L., Mooser, V., Morris, A. D., Palmer, C. N., Perry, J. R., Frayling, T. M., Whittaker, J. C., and Waterworth, D. M. (2012). Deep resequencing unveils genetic architecture of adipoq and identifies a novel low-frequency variant strongly associated with adiponectin variation. *Diabetes*, 61(5):1297–1301.

[147] Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

[148] Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82 – 93.

[149] Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics*, 45(3):pp. 939–955.

[150] Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1):pp. 175–188.

[151] Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

[152] Xu, J. and Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 50:375–387.

[153] Zeng, D. (2004). Estimating marginal survival function by adjusting for dependent censoring using many covariates. *The Annals of Statistics*, 32(4):pp. 1533–1555.

[154] Zeng, D. and Cai, J. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *The Annals of Statistics*, 33(5):2132–2163.

[155] Zhang, P. (2003). Multiple imputation: Theory and method. *International Statistical Review*, 71(3):581–592.

[156] Zhou, H. and Pepe, M. S. (1995). Auxiliary covariate data in failure time regression. *Biometrika*, 82(1):pp. 139–149.

[157] Zou, H., Hastie, T., and Tibshirani, R. (2004). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15:2006.