



Published in final edited form as:

J Pain. 2013 December ; 14(12 0): T102–T115. doi:10.1016/j.jpain.2013.09.003.

Multivariable modeling of phenotypic risk factors for first-onset TMD: the OPPERA prospective cohort study

Eric Bair^{1,2,3}, Richard Ohrbach⁴, Roger B. Fillingim⁵, Joel D. Greenspan⁶, Ronald Dubner⁶, Luda Diatchenko¹, Erika Helgeson³, Charles Knott⁷, William Maixner¹, and Gary D. Slade^{1,8,9}

¹ Regional Center for Neurosensory Disorders, University of North Carolina at Chapel Hill

² Department of Endodontics, University of North Carolina at Chapel Hill

³ Department of Biostatistics, University of North Carolina at Chapel Hill

⁴ Department of Oral Diagnostic Sciences, University at Buffalo

⁵ University of Florida, College of Dentistry, and Pain Research and Intervention Center of Excellence

⁶ Department of Neural and Pain Sciences, and Brotman Facial Pain Center, University of Maryland School of Dentistry

⁷ Battelle Memorial Institute, Center for Analytics and Public Health

⁸ Department of Dental Ecology, University of North Carolina at Chapel Hill

⁹ Department of Epidemiology, University of North Carolina at Chapel Hill

Abstract

Incidence of temporomandibular disorders (TMD) was predicted with multivariable models that used putative risk factors collected from initially TMD-free individuals in the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study. The 202 baseline risk factors included sociodemographic and clinical characteristics, measures of general health status, experimental pain sensitivity, autonomic function, and psychological distress. Study participants (n=2,737) were then followed prospectively for a median of 2.8 years to ascertain cases of first-onset TMD. Lasso regression and random forest models were used to predict incidence of first-onset TMD using all of the aforementioned measures. Variable importance scores identified the most important risk factors, and their relationship with TMD incidence was illustrated graphically

© 2013 The American Pain Society. Published by Elsevier Inc. All rights reserved.

Please address all correspondence to: Eric Bair CB #7450 School of Dentistry Chapel Hill, NC 27599-7450 Phone: (919) 537-3276 Fax: (919) 966-5339 ebair@email.unc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Disclosures

This work was supported by NIH grant U01DE017018. Roger Fillingim and Gary Slade are consultants and equity stock holders, and William Maixner and Dr. Luda Diatchenko are cofounders and equity stock holders in Algyonics, Inc., a company providing research services in personalized pain medication and diagnostics. Other authors declare no competing interests.

using partial dependence plots. Two of the most important risk factors for elevated TMD incidence were greater numbers of comorbid pain conditions and greater extent of non-specific orofacial symptoms. Other important baseline risk factors were pre-existing bodily pain, heightened somatic awareness, and greater extent of pain in response to examiners' palpation of the head, neck and body. Several demographic variables persisted as risk factors even after adjusting for other OPPERA variables, suggesting that environmental variables not measured in OPPERA may also contribute to first-onset TMD.

Keywords

chronic pain; multivariable analysis; data mining; OPPERA; temporomandibular disorders

Introduction

The primary objective of the OPPERA study was to identify potential risk factors for first-onset TMD. The study was motivated by a heuristic model (Supplementary Figure 1) postulating that heightened sensitivity to pain and psychological distress contribute to first-onset TMD.^{9, 27} These two main risk factors are in turn influenced by a variety of genetic and environmental variables.⁹ Thus, OPPERA collected data from each study participant on a wide variety of measures of experimental pain sensitivity and psychological distress as well as data on sociodemographic characteristics, autonomic function, clinical characteristics, and general health status.

As is common in epidemiologic studies of TMD,^{4, 22} the first stage of analysis examined univariate associations to test which individual variables predicted TMD onset.^{12, 14, 31, 34, 36} However, some questions of interest cannot be easily answered using univariate models. For example, one may wish to determine which risk factors for first-onset TMD are most important, that is, which variables provide the most information about the risk of developing first-onset TMD. One may also wish to evaluate the association between a given variable and first-onset TMD after adjusting for the effects of other variables. For example, we observed that race and lifetime U.S. residence were associated with first-onset TMD.³⁶ It seems unlikely that race or duration of U.S. residence directly causes orofacial pain, so the question arises whether these differences can be explained by other putative risk factors.

The present study seeks to answer these questions using two multivariable statistical methods. The first method uses a linear model for time-to-event data to predict the likelihood that a given study participant will develop first-onset TMD based on a subset of the variables that is selected to maximize the predictive accuracy of the model. The variables selected by such a procedure represent one possible set of the "most important variables," and the coefficients in such a model can be used to estimate the effect of each selected variable after adjusting for the other selected variables.

Despite the attractive simplicity of this approach, it has certain shortcomings. Any variable that is not selected by the procedure is not included in the model. Thus, there is no way to estimate the relative importance of variables that are not selected. More importantly, the

estimated effect sizes of each variable do not account for the effect of variables that are not selected by the model. Thus, if an important confounder of a given variable is not selected, the model may overestimate the importance of this variable.

Thus, a second stage of the analysis used a different multivariable approach, random forest modeling, to analyze potential contributions of all variables, not merely the variables selected by the earlier linear model. Random forest modeling represents a machine learning technique based on a series of decision tree models.²¹ Decision trees predict outcomes by recursively partitioning predictor variables, and these trees are superior to linear regression-based models in identifying non-linear effects and handling large numbers of correlated predictors.¹⁶ In recent years, random forests have been increasingly applied to classification problems in biomedical research, including predicting several pain related outcomes.^{19, 33, 42} This novel method of data mining was used to achieve two goals: a) to identify the most important risk factors for first-onset TMD; and b) to generate plots depicting adjusted association between each variable and TMD incidence, with adjustment for the effects of other variables and with latitude in generating the plots that permitted departure from a straight-line association.

Methods

Recruitment, Eligibility Criteria, and Enrollment

This paper reports findings from the OPPERA prospective cohort study of 2,737 people who were enrolled in 2006-08 and followed for a median of 2.8 years, during which time 260 people developed temporomandibular disorder (TMD). When they were enrolled, the sample of community-based volunteers at four study U.S. sites was aged 18-44 years and did not have painful TMD when examined using OPPERA's adaptation of a restricted set of Research Diagnostic Criteria for TMD (RDC/TMD).¹⁰ At enrollment, study participants also completed questionnaires, autonomic function was measured, sensitivity to sensory stimuli was evaluated, and a blood sample was collected for genotyping.

At three-monthly intervals after enrollment, study participants were asked to complete a screening questionnaire (Quarterly Health Update or QHU) that asked about TMD pain symptoms. Those reporting symptoms were invited to study clinics for a follow-up examination that determined presence or absence of painful TMD using the same criteria used at baseline. Specifically, the 260 incident cases satisfied two criteria for TMD: (1) symptoms of orofacial pain reported for ≥ 5 days/month; and (2) examiner findings of TMD myalgia, arthralgia, or both. An accompanying paper⁵ describes methods used to standardize examiners across study sites. Inter-examiner reliability between study site examiners and the OPPERA reference examiner was evaluated annually, yielding kappa statistics that ranged from 0.82 to 1.00, signifying excellent inter-examiner reliability. The paper⁵ also reports that QHU questionnaires had kappa values of 0.83 (95%CL = 0.72, 0.95) for test-retest reliability, indicating excellent agreement.

Institutional review boards at each study site approved study procedures and participants provided signed, informed consent. Full details of enrollment and follow-up are provided elsewhere in this volume.⁵

Lasso Regression Modeling

Lasso regression⁴¹ is a multivariable regression method that is useful for predicting an outcome measure in the presence of a large number of correlated predictor variables. Conventional least squares regression models minimize the sum of the squared differences between the predicted and actual values of the outcome variable. Lasso modifies this criterion by penalizing models that have too many large coefficients under the assumption that models with too many large coefficients are likely to be overfit. It can be shown¹⁶ that lasso models have lower variance than conventional least squares models thereby mitigating the effect of multicollinearity. Moreover, the lasso model selection criterion forces some of the model coefficients to be 0, so lasso also performs variable selection. The amount by which large coefficients are penalized is controlled by a tuning parameter. Larger values of the tuning parameter result in a larger penalty for the coefficients, resulting in fewer nonzero coefficients. The optimal tuning parameter can be chosen using cross-validation. See the appendix for a more detailed description of lasso regression.

This paper focuses on baseline measurements from the following six risk domains: 1) sociodemographic variables, 2) measures of experimental pain sensitivity, 3) measures of autonomic function, 4) measures of psychological functioning, 5) measures of general health status, 6) clinical orofacial characteristics. A seventh grouping of variables from all domains was designated the “cross-domain” set of variables. The variables included in the cross-domain model (and the OPPERA instruments that evaluated each variable) are listed in Supplementary Table 1. Data collection methods used to create the variables have been published previously^{11, 13, 30, 35} and in accompanying papers in this volume.^{12, 14, 31, 34, 36}

A lasso regression model was calculated using all of the variables collected in the six domains of interest. Any variable with more than 150 missing values was excluded from the analysis. This threshold was chosen because only a small number of predictor variables had more than 150 missing values, and including these variables in the model created computational problems. Missing values of the remaining variables were imputed using the EM algorithm as described previously.^{12, 13, 35} A total of 202 variables collected from 2,734 participants were included in the analysis. (Three participants were excluded due to excessive missing data.) Dummy variables were created for categorical variables and all variables (including these dummy variables) were normalized to have mean 0 and standard deviation 1 prior to fitting the model.

The outcome variable was the time from enrollment to when the participant was classified as meeting the criteria for first-onset TMD. However, this analysis requires one to adjust for the fact that many study participants did not develop first-onset TMD. A set of statistical tools known as survival analysis methods can be used for this type of data. In survival analysis, the outcome is the time until an event occurs (development of first-onset TMD in the present study). If no event occurs before the end of the follow up period (or if a given individual drops out of the study prematurely), the observation is said to be censored. Lasso regression can be applied when the outcome is a censored survival time.⁴⁰

The tuning parameter for the lasso model was selected using cross-validation. After the optimal tuning parameter was identified, the nonzero lasso regression coefficients were

calculated for the optimal model. Each regression coefficient was then converted to a hazard ratio (HR). For continuous predictor variables, only the standardized HR (the HR after normalizing the predictor variable to have mean 0 and standard deviation 1) was reported. Standardized HR's allow for the comparison of the effect sizes of variables that may be measured using different scales. Both standardized and unstandardized HR's were reported for categorical variables. (Standardized HR's for categorical variables cannot be easily interpreted. However, they are still useful for comparing the relative effect sizes of the variables in the model.) The lasso model was calculated using version 1.7.3 of the “glmnet” R package in R version 2.13.1.

Random Forest Modeling

A second form of multivariable modeling was performed using random forests, which are based on a series of decision tree models. A decision tree predicts an outcome by recursively partitioning the set of predictor variables producing results that can be visualized as a tree diagram.⁷ Compared to conventional Cox models, decision trees can readily detect nonlinear effects and interactions, and they are robust against missing values. However, decision trees typically have high variance, resulting in inaccurate predictions.¹⁶ Random forests attempt to capture the advantages of trees while reducing their variance by averaging over a series of decision trees. Each decision tree is fit using a subset of the data. The final predictions are obtained by averaging over 1,000 decision trees. Only a subset of the predictors is used in each tree to reduce the correlation between the trees.^{6, 16} All variables are included in the final model, however, since it consists of the average of 1,000 such trees. Thus, unlike a simple decision tree, the output of a random forest cannot be visualized as a tree diagram. For a description of how random forests can be applied to censored survival data, see Ishwaran et al. (2008).²¹ See the appendix for a more detailed description of the random forest methodology.

A random forest model was calculated using all of the variables collected in the six domains of interest described previously. Unless otherwise stated, OPPERA study site was used in each model. Any variable with more than 150 missing values was excluded from the analysis. Missing values of the remaining variables were imputed by adaptive tree imputation,²¹ with the exception of the pain sensitivity, autonomic, and psychosocial variables, which were imputed using the EM algorithm imputed as described previously.^{11, 13, 35} Censoring indicators were also imputed for 318 participants who were identified as possible cases of first-onset TMD based on their responses to the QHU but never returned to the clinic to undergo an RDC/TMD examination for definitive classification of first-onset TMD. (Note that this imputation was performed using adaptive tree imputation, and therefore was different from the multiple imputation method described elsewhere in this volume⁵ that imputed outcomes from the same 318 participants.) One consequence is that estimated incidence rates produced by the random forest models may differ slightly from the estimated incidence rates reported elsewhere. As was the case for the lasso model, a total of 202 variables collected from 2,734 participants were included in the analysis.

The first objective for fitting the random forest model was to compare the predictive accuracy of the variables in each risk domain for predicting first-onset TMD. The predictive accuracy of each random forest model was evaluated by calculating Harrell's concordance index (HCI).¹⁵ The index is a generalization of the area under the receiver operating characteristic (ROC) curve applied to censored survival data. A HCI of 1.0 corresponds to perfect predictive accuracy; whereas a HCI of 0.5 means that the model is no better than random guessing. In studies of diagnostic tests, descriptors of “rather low”, “useful” and “rather high” accuracy of prediction have been proposed³⁹ for area-under-ROC-curve statistics at thresholds of <0.7, 0.7 to 0.9, and >0.9, respectively, although the thresholds clearly are arbitrary. Six HCI's were calculated, one from each random forest model fit to the respective domains of interest. In addition, the HCI was calculated for the model that included all six domains.

The second objective for fitting the random forest model was to identify the most important variables for predicting first-onset TMD. This was assessed by calculating the variable importance score (VIS), which estimates the decrease in the predictive accuracy of the model when the variable is measured incorrectly. When the variable is an important predictor, the decrease will be large. By convention, the most important variable is scaled to have a VIS of 100, and all other VIS's have lower values. A VIS of zero signifies that the predictive accuracy is not decreased when the variable is measured with error, and a negative VIS indicates that predictive accuracy increases when the variable is measured incorrectly. See the appendix for a more detailed description of how the VIS's are calculated.

The third objective for fitting the random forest model was to estimate the association between each variable and first-onset TMD after adjusting for the effects of all other variables. For example, previous analysis revealed an association between race and lifetime U.S. residence and first-onset TMD.³⁶ It seems unlikely that lifetime U.S. residence directly causes TMD. Thus, one may wish to determine if this association can be explained by other variables measured in OPPERA. Perhaps lifetime U.S. residents have more comorbid pain conditions, more pre-existing pain, or greater somatic awareness (or greater levels of any other potential risk factor measured in OPPERA). If this is the case, then the association between lifetime U.S. residence and first-onset TMD will disappear after adjusting for these other variables. However, if this association remains after adjusting for other variables, this suggests that other environmental variables not measured in OPPERA that may also influence the risk of developing first-onset TMD.

The association between each variable and first-onset TMD after adjusting for the effects of other variables was assessed by estimating the expected rate of first-onset TMD that would be observed at several values of the variable after averaging over the values of all other variables in the model, and the results were plotted. Partial dependence plots were estimated at up to 25 values of continuous predictor variables and a loess smoother (and associated 95% confidence interval) was calculated to help visualize the association. For a more detailed description of loess smoothing, see the appendix or Loader (1999).²³ For categorical variables, the estimated incidence rate was calculated for each participant and box plots of the estimated incidence rates were plotted separately for each category. Partial dependence plots were generated for the four variables with the highest VIS's in each of the

six OPPERA domains. They were also calculated for each variable with a nonzero lasso coefficient as well as a few other variables of interest. See the appendix for a more detailed description of how these partial dependence plots were calculated. The random forest models were fitted using version 3.6.3 of the “randomSurvivalForest” R package in R version 2.13.1.^{20, 21}

Results

Lasso Regression

The HR's estimated from the nonzero coefficients in the lasso regression model are shown in Table 1. (All variables not included in Table 1 had a regression coefficient of 0.) The three strongest predictors of first-onset TMD are the somatization subscale of the SCL-90R from the psychosocial domain, the count of comorbid conditions from the health status domain, and the count of non-specific orofacial symptoms from the clinical domain. Overall there were six health status variables, five clinical variables, two psychosocial variables, one sociodemographic and one autonomic variable with nonzero coefficients. The dummy variables for the Florida study site, smoking, and lifetime U.S. residence also had nonzero coefficients. Note that all other sociodemographic variables (including age, race, and gender) had coefficients of 0. It is also noteworthy that one autonomic variable (namely average diastolic blood pressure during the orthostatic challenge) had a nonzero coefficient despite the fact that this variable was not associated with first-onset TMD in the univariate analysis.¹⁴

Comparison of the Random Forest Models

HCI values for each of the seven random forest models ranged from 0.65 for the autonomic domain to 0.75 for the health status domain (Table 2). For the cross-domain model, the HCI was 0.74. All these values are close to the threshold that has been proposed to distinguish between “rather low” and “useful” accuracy of prediction. However, there is no formal statistical test to distinguish between the HCI values shown in Table 2.

Variable Importance Scores

The 30 most important predictors of first-onset TMD based on VIS from the random forest model are shown in Table 3. The four most important predictors had VIS values ranging from 100 to 80: 1) the number of comorbid conditions from the health status domain, 2) the number of non-specific orofacial symptoms from the clinical domain, 3) OPPERA study site and 4) the SF-12 bodily pain score from the health status domain. There were six other predictors with VIS values ranging from 70 to 40, and all remaining predictors had VIS of 33.3 or lower.

Table 4 shows the five variables with the largest VIS within each of the six OPPERA domains (as well as each variable's overall VIS rank). All five of these variables had a VIS of 10.0 or greater (placing them among the 30 most important variables overall) for the health status, clinical, sociodemographic, and autonomic domains. In contrast, none of the pain sensitivity variables had a VIS of 10.0 or greater, and only two of the top five

psychosocial variables had a VIS of 10.0 or greater (although four of the top five psychosocial variables had a VIS of 9.7 or greater).

The two most important sociodemographic predictors were age (which ranged from 18-44 years in this study) and marital status, with lesser contribution to prediction from race, and lifetime U.S. residence. Gender was not an important predictor of first-onset TMD (VIS=-8.3). Among the top five health status variables, three of the strongest effects were seen for the number of comorbid conditions, the SF-12 bodily pain score, and the SF-12 general health score. The remaining health status variables had noticeably lower importance scores. Four of the top five clinical predictors had a VIS of 50.0 or greater, including the non-specific orofacial symptoms, oral parafunctions, reported limitation of mouth opening, and the number of craniofacial palpation tender points. A history of orthodontic treatment also had a relatively high VIS (VIS=29.3).

In the psychosocial domain, the most important predictor was the PILL score of somatic awareness (VIS=42.4). The remaining psychosocial variables had much smaller importance scores. In the experimental pain sensitivity domain, all variables had importance scores of 5.8 or lower. Interestingly, each of the five most important autonomic predictors, with importance scores ranging from 10.8 to 19.1, were measures recorded during the Stroop stress procedure: total power, mean arterial blood pressure and heart rate. It is also noteworthy that all five of the strongest autonomic predictors had higher importance scores than any of the pain sensitivity variables and any of the psychosocial variables except PILL.

Partial Dependence Plots

The partial dependence plot for OPPERA study site is shown in Supplementary Figure 2. A boxplot of the estimated distribution of the TMD incidence rate (after adjusting for other OPPERA variables) is calculated for each study site. The rate of TMD incidence shows substantial variation between study sites, with the Buffalo, NY and the Gainesville, FL study sites showing higher rates of TMD incidence than the Chapel Hill, NC and Baltimore, MD sites.

The partial dependence plots for age, marital status, race, and lifetime U.S. residence are shown in Figure 1, and the partial dependence plot for gender is shown in Supplementary Figure 3. For continuous variables, the partial dependence plots estimate the TMD incidence rate for a series of possible values in the range of the variable after adjusting for other OPPERA variables. Consistent with findings reported elsewhere in this volume,³⁶ greater age was associated with first-onset TMD, and there was no significant association between gender and first-onset TMD. However, partial dependence boxplots for race differed from the univariate findings. While African-Americans had a higher rate of first-onset TMD compared to Whites, Asians did not have a lower incidence rate. Similarly, marital status, which was not associated with first-onset TMD univariate analysis, was strongly associated with first-onset TMD in the partial dependence plots, with married or previously married individuals showing higher incidence rates.

The partial dependence plots for the health status variables are shown in Figure 2 and Supplementary Figure 4. Consistent with the findings from Cox models reported

elsewhere,³⁴ a greater number of comorbid conditions, a greater number of headaches, and a greater number of IBS symptoms were associated with first-onset TMD. Likewise, lower scores on the SF-12 bodily pain and general health scales (corresponding to higher pain levels and poorer levels of general health) and a higher PSQI score were associated with first-onset TMD. Finally, both current and former smokers had higher rates of first-onset TMD, as observed in the univariate Cox models.³⁴ (See Supplementary Figure 4.)

The partial dependence plots for the clinical variables are shown in Figure 3 and Supplementary Figure 5. The results are nearly identical to the results reported in previous Cox models: reported limitation of mouth opening was associated with first-onset TMD, as were a greater number of non-specific orofacial symptoms, a greater number of palpation sites with pain, and oral parafunctions. One interesting finding was that a history of orthodontic procedures was associated with lower TMD incidence. Although a similar trend was observed in the univariate analysis, that association was not statistically significant.³¹

The partial dependence plots for the psychosocial variables are shown in Figure 4 and Supplementary Figures 6 and 7. As seen in the univariate analysis, the PILL score (and SCL 90R score) of somatic awareness had a positive association with first-onset TMD, whereas effects were less apparent for other psychosocial variables. Likewise, Figure 5 and Supplementary Figure 8 show that the effects of pain sensitivity variables are generally consistent with the univariate analysis. Specifically, greater sensitivity to both pressure and thermal stimuli was associated with greater rates of first-onset TMD, although the effects were weak.

Figure 6 and Supplementary Figures 9 and 10 show partial dependence plots for the autonomic variables, providing an interesting contrast with the univariate results. In the univariate analysis, no significant relationship was observed between any of the autonomic measures and first-onset TMD. In this multivariable analysis, however, higher diastolic blood pressure during the orthostatic challenge, higher evoked mean arterial pressure during both the color and pain-affective STROOP procedures, and lower HRV all show an association with first-onset TMD.

Discussion

This analysis evaluated over 200 variables measured when participants were enrolled into the OPPERA prospective cohort, with the goal to identify variables that best predicted first-onset TMD. The study purposefully measured a large number of variables, recognizing that TMD is influenced by a multitude of factors, most of which have several facets that warrant evaluation. Their influences on TMD incidence were hypothesized *a priori* in a heuristic model developed for this study (Supplementary Figure 1).^{9, 27} The present analysis used lasso regression and random forests, novel multivariable analysis methods that are well-suited to deal with problems created by the number and density of data collected in this cohort study. The findings support the heuristic model by demonstrating a prominent contribution of psychological distress, particularly somatic symptoms. Pain amplification and autonomic function had smaller, yet measurable, effects that supported the basic domains of the proposed heuristic model. In addition, the findings reveal pronounced effects

on TMD incidence of two domains that are not explicitly depicted in the heuristic model: clinical orofacial characteristics and general health status.

One of the strongest predictors of first-onset TMD in both models was the number of comorbid conditions reported by the participant. The comorbid conditions included some painful conditions such as fibromyalgia and lower back pain, although it included other conditions that are not primarily painful (e.g. depression and sleep apnea). One explanation for the predominance of these variables relates to the model building process. Specifically, these methods select from among the many predictor variables by singling out those that best predict TMD incidence with the least measurement “noise.” This means that, given the choice between two variables that predict a proportion of total TMD incidence with some level of noise and a third variable that predicts the same proportion of total TMD incidence with less noise, the method will select the third variable. By this means, a seemingly heterogeneous measure, such as the number of comorbid conditions, can be selected by the model if it effectively captures variability in TMD incidence with less noise than several more specific measures that are strongly associated with comorbid conditions. In this instance, there is good evidence that people with numerous comorbid conditions have correspondingly high levels of somatic awareness, muscle tenderness, fatigue, and “unexplained” symptoms.^{1-3, 29} Thus, it is not surprising that this measure has the highest importance score of any variable collected in OPPERA. Likewise the bodily pain score of the SF-12, which was also a strong predictor in both models, probably captures many of these other variables as well.

Another interesting finding was a small but noticeable association between several autonomic variables and first-onset TMD. In the OPPERA case-control study, we observed that chronic TMD was associated with higher resting and stress evoked heart rates and lower HRV,²⁸ and similar findings have been observed among patients with comorbid conditions such as fibromyalgia.^{8, 32, 37} This is consistent with our other studies²⁴⁻²⁶ showing that dysregulation of autonomic modulatory systems may contribute to risk of TMD and related conditions. It is unclear why the multivariable models detected an association between the autonomic variables and first-onset TMD whereas no association was observed in the univariate analysis. One possibility is that there is an interaction between the autonomic variables and other OPPERA variables. A similar phenomenon was observed for some other, seeming unrelated variables, including marital status. Further investigation will be needed to explain these incongruous findings.

The random forest models also illustrate that much is still unknown about the etiology of TMD. We suggested earlier that the sociodemographic measures collected in OPPERA most likely do not contribute directly to first-onset TMD but rather are associated with other risk factors for TMD, and we hypothesized that many demographic differences in TMD onset could be explained once we account for the effect of the other OPPERA variables.³⁶ In some cases this hypothesis is supported. After accounting for these differences, the effect of lifetime U.S. residence on first-onset TMD largely disappears (Figure 1). However, even after accounting for other OPPERA variables there are still strong associations between first-onset TMD and both race and study site. Similarly, the model identified an association between first-onset TMD and marital status and history of orthodontic treatment. Given that

none of these variables are likely to directly cause first-onset TMD, these results suggest that there are other cultural or environmental variables not measured in OPPERA that contribute to first-onset TMD. Indeed, the cross-domain random forest model had only a slightly higher HCI than the random forest model for the sociodemographic variables, suggesting that these unknown cultural/environmental variables may be responsible for a large proportion of the variability in TMD incidence.

It is also interesting to note that the standardized HR's associated with the variables in the lasso model were noticeably attenuated compared to the corresponding univariate HR's. For example, the SCL 90R somatization score had a standardized HR of 1.18 in the lasso model (compared to 1.38 in the univariate analysis)¹² and the PSQI sleep score had a standardized HR of 1.07 (compared to 1.32 in the univariate analysis).³⁴ Although standardized HR's cannot be computed for the random forests models, a similar result was evident in the partial dependence plots. This suggests that the etiology of first-onset TMD is strongly multifactorial and that any single factor is likely confounded to some degree (although not completely) by other factors. Although many of the measures collected in OPPERA are correlated with one another, one cannot identify a single variable that captures all the information in the individual variables. Each of the variables provides a unique contribution to the risk of developing TMD although the independent contribution of any individual variable is small.

In general, the results of the lasso model were similar to the results of the random forest model. Both models found that a history of comorbid conditions, bodily pain, non-specific orofacial symptoms, and somatic awareness were among the most important predictors of first-onset TMD. However, some minor differences between the two models were also apparent. For example, the random forest model found a strong association between marital status and first-onset TMD whereas marital status was not selected by the lasso model. Given that marital status also was not significant in the univariate analysis,³⁶ this suggests that marital status is only associated with first-onset TMD via an interaction with some other variables that were not included in the lasso model. The random forests can detect this association since it adjusts for all 202 variables rather than the subset selected by lasso.

Other differences between the two models can be explained by the fact that many OPPERA variables are strongly correlated with one another. For example, the number of palpation sites with pain in the right masseter had a high VIS in the random forest model whereas the lasso model selected the count of palpation sites with pain in the right and left temporalis. Similarly, both models include variables measuring somatic awareness and autonomic function, but different variables are selected by each model. The most likely explanation is that the variables are approximately interchangeable: one could substitute the PILL somatization score for the SCL 90R somatization score without noticeably changing the model. Thus, one should be cautious about assuming that a variable with a low VIS or a variable not selected by the lasso model is “not important.” Such variables may simply be strongly correlated with another variable with a greater lasso coefficient or VIS. Indeed, one potential shortcoming of both lasso and random forests is that there is no simple statistical test to determine if a variable is “significantly” associated with first-onset TMD after adjusting for other variables (nor is there a simple way to calculate confidence intervals for

the lasso coefficients or VIS's). Determining which variables are truly superfluous and which variables are merely correlated with other important variables is a difficult problem and an area for future research.

Another methodological consideration is that lasso and random forests are optimized to discriminate between people who develop first-onset TMD and people who do not. An optimal statistical model for discrimination can readily overlook etiologic factors that contribute more subtly to an individual's degree of risk. For example, family history is often used to predict risk of cardiovascular disease, in part because family history often captures aspects of inherited disposition and environmental circumstances that are relevant to the disease. Hence, a statistical procedure to optimize discrimination between diseased and healthy people might select a simple measure of family history rather than a more complex set of measures relating to inheritance and environment. In the case of first-onset TMD, as noted earlier, it is conceivable that comorbid conditions, the single most-important predictor, reflect a range of biological processes that have emerged over time to manifest as one or more comorbid conditions. Complete identification of such factors may require further analysis and research.

Another important area for future research is the development of a model that could be used to predict an individual's risk of developing first-onset TMD. Although the models described in the present manuscript could be used for this purpose, in practice they are unsuitable for several reasons. For such a model to be useful in a clinical situation, it would need to be based on a small number of variables that can be easily measured and applied to a patient. The random forest model is unsuitable for that purpose since it would require measuring all 202 variables considered in the present study. Although the lasso model uses a smaller number of variables, ease of measuring such variables was not considered when fitting the model. It would also need further validation to verify that the predictive accuracy of the model is satisfactory³⁸ and that applying the model in practice would not be unsafe or excessively costly.^{17, 18} Instead, the primary goal of these models is to identify a set of variables that are important predictors of the overall rate of TMD in this group of study participants. The value of these models therefore is in understanding predictors of the overall rate of TMD, not in predicting an individual's risk of TMD.

These limitations notwithstanding, the current findings identify several environmental, clinical, psychosocial, autonomic and neurosensory measures that predict TMD incidence, providing support for the OPPERA heuristic model. The current approach has demonstrated the power of lasso regression and random forest modeling in evaluating a seemingly dense array of measures across multiple domains. This represents a novel application of these methods in pain research where there is a pressing need to evaluate multiple, interacting risk factors in order to understand complex pain conditions. The findings demonstrate that an extensive array of risk factors contribute to the development of TMD. In revealing prominent contributions from clinical and health status domains that have not been clearly depicted in existing models of TMD etiology, the results offer new directions for future studies of this complex, biopsychosocial illness.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The OPPERA program acknowledges resources specifically provided for this project by the respective host universities: University at Buffalo, University of Florida, University of Maryland-Baltimore, and University of North Carolina-Chapel Hill. The authors would like to thank the OPPERA Research Staff for their invaluable contributions to this work. In addition, we express our gratitude to the research participants who have devoted time and effort in support of this research.

References

1. Aaron LA, Burke MM, Buchwald D. Overlapping conditions among patients with chronic fatigue syndrome, fibromyalgia, and temporomandibular disorder. *Arch Intern Med.* 2000; 160(2):221–227. [PubMed: 10647761]
2. Aaron LA, Buchwald D. A review of the evidence for overlap among unexplained clinical conditions. *Ann Intern Med.* 2001; 134(9 Pt 2):868–881. [PubMed: 11346323]
3. Aaron LA, Buchwald D: Chronic diffuse musculoskeletal pain, fibromyalgia and co-morbid unexplained clinical conditions. *Best Pract Res Clin Rheumatol.* 2003; 17(4):563–574. [PubMed: 12849712]
4. Aggarwal VR, Macfarlane GJ, Farragher TM, McBeth J. Risk factors for onset of chronic oro- facial pain--results of the North Cheshire oro-facial pain prospective population study. *Pain.* 2010; 149(2): 354–359. [PubMed: 20304556]
5. Bair E, Brownstein NC, Ohrbach R, Greenspan JD, Dubner R, Fillingim RB, Maixner W, Smith SB, Diatchenko L, Gonzalez Y, Gordon SM, Lim PF, Ribeiro-Dasilva M, Dampier D, Knott C, Slade GD. Study protocol, sample characteristics, and loss to follow-up: the OPPERA prospective cohort study. *J Pain.* 2013 Dec; 14(12 Suppl):T2–19. [PubMed: 24275220]
6. Breiman LL. Random forests. *Mach Learning.* 2001; 45(1):5–32.
7. Breiman, L.; Friedman, JH.; Stone, CJ.; Olshen, RA. Classification and regression trees. Chapman & Hall : International Thomson Publishing; New York, N.Y.: 1984. p. 358
8. Cohen H, Neumann L, Shore M, Amir M, Cassuto Y, Buskila D. Autonomic dysfunction in patients with fibromyalgia: application of power spectral analysis of heart rate variability. *Semin Arthritis Rheum.* 2000; 29(4):217–227. [PubMed: 10707990]
9. Diatchenko L, Nackley AG, Slade GD, Fillingim RB, Maixner W. Idiopathic pain disorders--pathways of vulnerability. *Pain.* 2006; 123(3):226–230. [PubMed: 16777329]
10. Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. *J Craniomandib Disord.* 1992; 6(4):301–355. [PubMed: 1298767]
11. Fillingim RB, Ohrbach R, Greenspan JD, Knott C, Dubner R, Bair E, Baraian C, Slade GD, Maixner W. Potential psychosocial risk factors for chronic TMD: descriptive data and empirically identified domains from the OPPERA case-control study. *J Pain.* 2011; 12(11 Suppl):T46–60. [PubMed: 22074752]
12. Fillingim RB, Ohrbach R, Greenspan JD, Knott C, Diatchenko L, Dubner R, Bair E, Baraian C, Mack N, Slade GD, Maixner W. Psychological factors associated with development of TMD: the OPPERA prospective cohort study. *J Pain.* 2013 Dec; 14(12 Suppl):T75–90. [PubMed: 24275225]
13. Greenspan JD, Slade GD, Bair E, Dubner R, Fillingim RB, Ohrbach R, Knott C, Mulkey F, Rothwell R, Maixner W. Pain sensitivity risk factors for chronic TMD: descriptive data and empirically identified domains from the OPPERA case control study. *J Pain.* 2011; 12(11 Suppl):T61–74. [PubMed: 22074753]
14. Greenspan JD, Slade GD, Bair E, Dubner R, Fillingim RB, Ohrbach R, Knott C, Diatchenko L, Liu Q, Maixner W. Pain sensitivity and autonomic factors associated with development of TMD: the OPPERA prospective cohort study. *J Pain.* 2013 Dec; 14(12 Suppl):T63–74. [PubMed: 24275224]

15. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. 1982; 247(18):2543–2546. [PubMed: 7069920]
16. Hastie, T.; Tibshirani, R.; Friedman, JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer; New York: 2009. p. 745
17. Helfand M, Buckley DI, Freeman M, Fu R, Rogers K, Fleming C, Humphrey LL. Emerging risk factors for coronary heart disease: a summary of systematic reviews conducted for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2009; 151(7):496–507. [PubMed: 19805772]
18. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MSV, Go AS, Harrell FE, Hong Y, Howard BV, Howard VJ, Hsue PY, Kramer CM, McConnell JP, Normand ST, O'Donnell CJ, Smith SC. Wilson PWF on behalf of the American Heart Association Expert Panel on Subclinical Atherosclerotic Diseases and Emerging Risk Factors and the Stroke Council: Criteria for Evaluation of Novel Markers of Cardiovascular Risk: A Scientific Statement From the American Heart Association. *Circulation*. 2009; 119(17):2408–2416. [PubMed: 19364974]
19. Hu YJ, Ku TH, Jan RH, Wang K, Tseng YC, Yang SF. Decision tree-based learning to predict patient controlled analgesia consumption and readjustment. *BMC Med Inform Decis Mak*. 2012; 12:131–6947-12-131. [PubMed: 23148492]
20. Ishwaran H, Kogalur UB. Random survival forests for R. *R News*. 2007; 7(2):25–31.
21. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. *The Annals of Applied Statistics*. 2008; 2(3):841–860.
22. Kino K, Sugisaki M, Ishikawa T, Shibuya T, Amagasa T, Miyaoka H. Preliminary psychologic survey of orofacial outpatients. Part 1: Predictors of anxiety or depression. *J Orofac Pain*. 2001; 15(3):235–244. [PubMed: 11575194]
23. Loader, C. *Local regression and likelihood*. Springer; New York: 1999.
24. Maixner W, Fillingim R, Booker D, Sigurdsson A. Sensitivity of patients with painful temporomandibular disorders to experimentally evoked pain. *Pain*. 1995; 63(3):341–351. [PubMed: 8719535]
25. Maixner, W.; Sigurdsson, A.; Fillingim, RB.; Lundeen, T.; Booker, D. Regulation of acute and chronic orofacial pain. In: Friction, JR.; Dubner, R., editors. *Orofacial Pain and Temporomandibular Disorders*. Raven Press, Ltd.; New York, NY: 1995. p. 85-102.
26. Maixner, W. Biopsychological and genetic risk factors for temporomandibular joint disorders and related conditions. In: Graven-Nielsen, T.; Arendt-Nielsen, L.; Mense, S., editors. *Fundamentals of Musculoskeletal Pain*. IASP Press; Seattle, WA: 2008. p. 263-279.
27. Maixner W, Diatchenko L, Dubner R, Fillingim RB, Greenspan JD, Knott C, Ohrbach R, Weir B, Slade GD. Orofacial Pain Prospective Evaluation and Risk Assessment study--the OPPERA study. *J Pain*. 2011; 12(11 Suppl):T4–11. e1–2. [PubMed: 22074751]
28. Maixner W, Greenspan JD, Dubner R, Bair E, Mulkey F, Miller V, Knott C, Slade GD, Ohrbach R, Diatchenko L, Fillingim RB. Potential autonomic risk factors for chronic TMD: descriptive data and empirically identified domains from the OPPERA case-control study. *J Pain*. 2011; 12(11 Suppl):T75–91. [PubMed: 22074754]
29. Nickel JC, Tripp DA, Pontari M, Moldwin R, Mayer R, Carr LK, Doggweiler R, Yang CC, Mishra N, Nordling J. Interstitial cystitis/painful bladder syndrome and associated medical conditions with an emphasis on irritable bowel syndrome, fibromyalgia and chronic fatigue syndrome. *J Urol*. 2010; 184(4):1358–1363. [PubMed: 20719340]
30. Ohrbach R, Fillingim RB, Mulkey F, Gonzalez Y, Gordon S, Gremillion H, Lim PF, Ribeiro-Dasilva M, Greenspan JD, Knott C, Maixner W, Slade G. Clinical findings and pain symptoms as potential risk factors for chronic TMD: descriptive data and empirically identified domains from the OPPERA case-control study. *J Pain*. 2011; 12(11 Suppl):T27–45. [PubMed: 22074750]
31. Ohrbach R, Bair E, Fillingim RB, Gonzalez Y, Gordon SM, Lim PF, Ribeiro-Dasilva M, Diatchenko L, Dubner R, Greenspan JD, Knott C, Maixner W, Smith SB, Slade GD. Clinical orofacial characteristics associated with risk of first-onset TMD: the OPPERA prospective cohort study. *J Pain*. 2013 Dec; 14(12 Suppl):T33–50. [PubMed: 24275222]
32. Reyes Del Paso GA, Garrido S, Pulgar A, Martin-Vazquez M, Duschek S. Aberrances in autonomic cardiovascular regulation in fibromyalgia syndrome and their relevance for clinical pain reports. *Psychosom Med*. 2010; 72(5):462–470. [PubMed: 20467004]

33. Riddle DL, Kong X, Jiranek WA. Two-year incidence and predictors of future knee arthroplasty in persons with symptomatic knee osteoarthritis: preliminary analysis of longitudinal data from the osteoarthritis initiative. *Knee*. 2009; 16(6):494–500. [PubMed: 19419874]
34. Sanders AE, Slade GD, Bair E, Fillingim RB, Knott C, Dubner R, Greenspan JD, Maixner W, Ohrbach R. General health status and incidence of first-onset temporomandibular disorder: the OPPERA prospective cohort study. *J Pain*. 2013 Dec; 14(12 Suppl):T51–62. [PubMed: 24275223]
35. Slade GD, Bair E, By K, Mulkey F, Baraian C, Rothwell R, Reynolds M, Miller V, Gonzalez Y, Gordon S, Ribeiro-Dasilva M, Lim PF, Greenspan JD, Dubner R, Fillingim RB, Diatchenko L, Maixner W, Dampier D, Knott C, Ohrbach R. Study methods, recruitment, sociodemographic findings, and demographic representativeness in the OPPERA study. *J Pain*. 2011; 12(11 Suppl):T12–26. [PubMed: 22074749]
36. Slade GD, Bair E, Greenspan JD, Dubner R, Fillingim RB, Diatchenko L, Maixner W, Knott C, Ohrbach R. Signs and symptoms of first-onset TMD and sociodemographic predictors of its development: the OPPERA prospective cohort study. *J Pain*. 2013 Dec; 14(12 Suppl):T20–32. [PubMed: 24275221]
37. Solberg Nes L, Carlson CR, Crofford LJ, de Leeuw R, Segerstrom SC. Self-regulatory deficits in fibromyalgia and temporomandibular disorders. *Pain*. 2010; 151(1):37–44. [PubMed: 20561749]
38. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010; 21(1):128–138. [PubMed: 20010215]
39. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988; 240(4857):1285–1293. [PubMed: 3287615]
40. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997; 16(4): 385–395. [PubMed: 9044528]
41. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996; 58(1):267–288.
42. Wolfe F, Clauw DJ, Fitzcharles MA, Goldenberg DL, Katz RS, Mease P, Russell AS, Russell IJ, Winfield JB, Yunus MB. The American College of Rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity. *Arthritis Care Res (Hoboken)*. 2010; 62(5):600–610. [PubMed: 20461783]

Perspective

Multivariable methods were used to identify the most important predictors of first-onset TMD in the OPPERA study. Important variables included comorbid pain conditions, pre-existing pain, and somatic awareness. Demographic characteristics, which probably reflect environmental variables not measured in OPPERA, also appear to play an important role in the etiology of TMD.

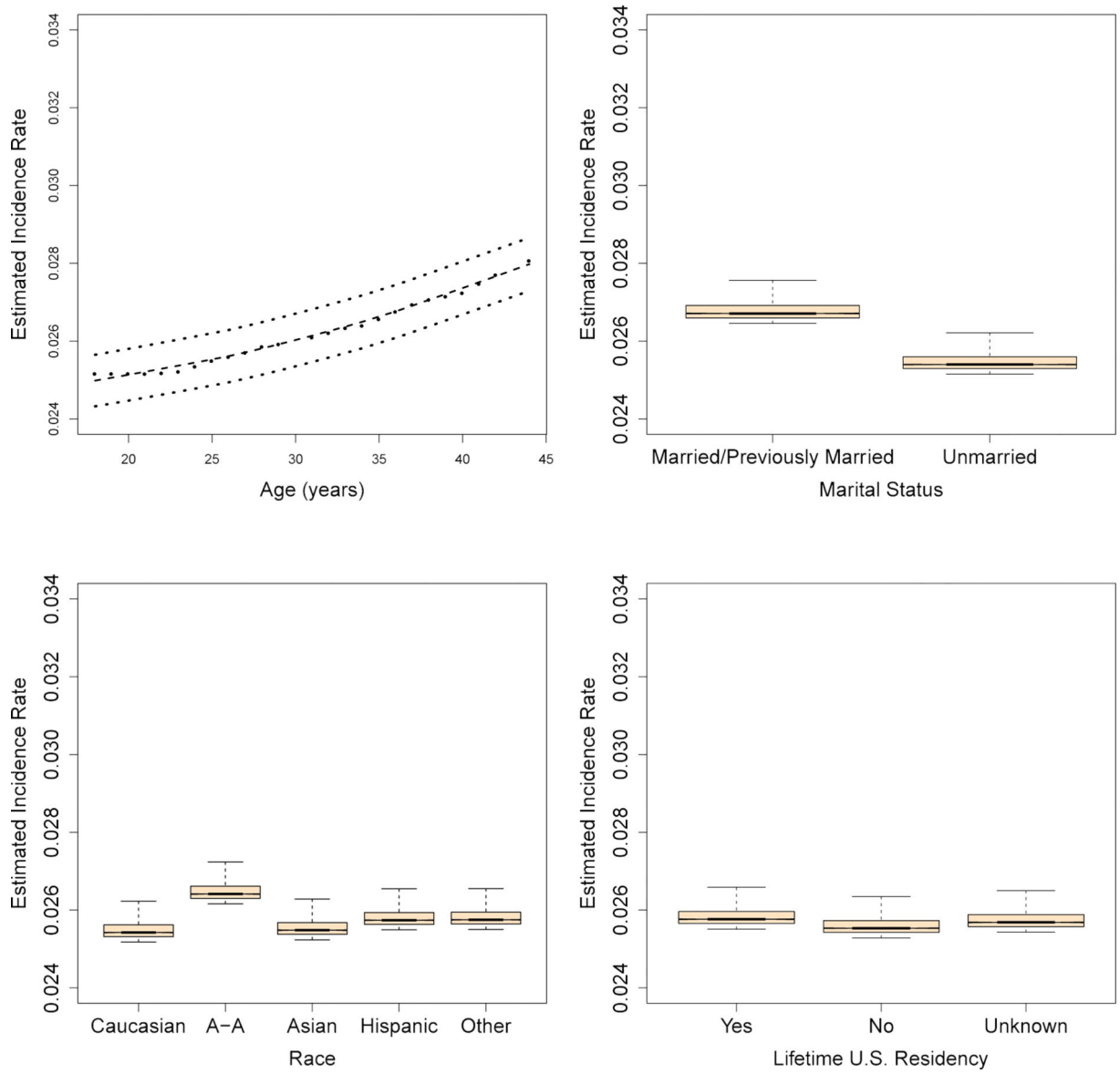


Figure 1. Partial dependence plots for selected sociodemographic variables, which show the estimated TMD incidence rate for several possible values of each variable after adjusting for all other OPPERA variables.

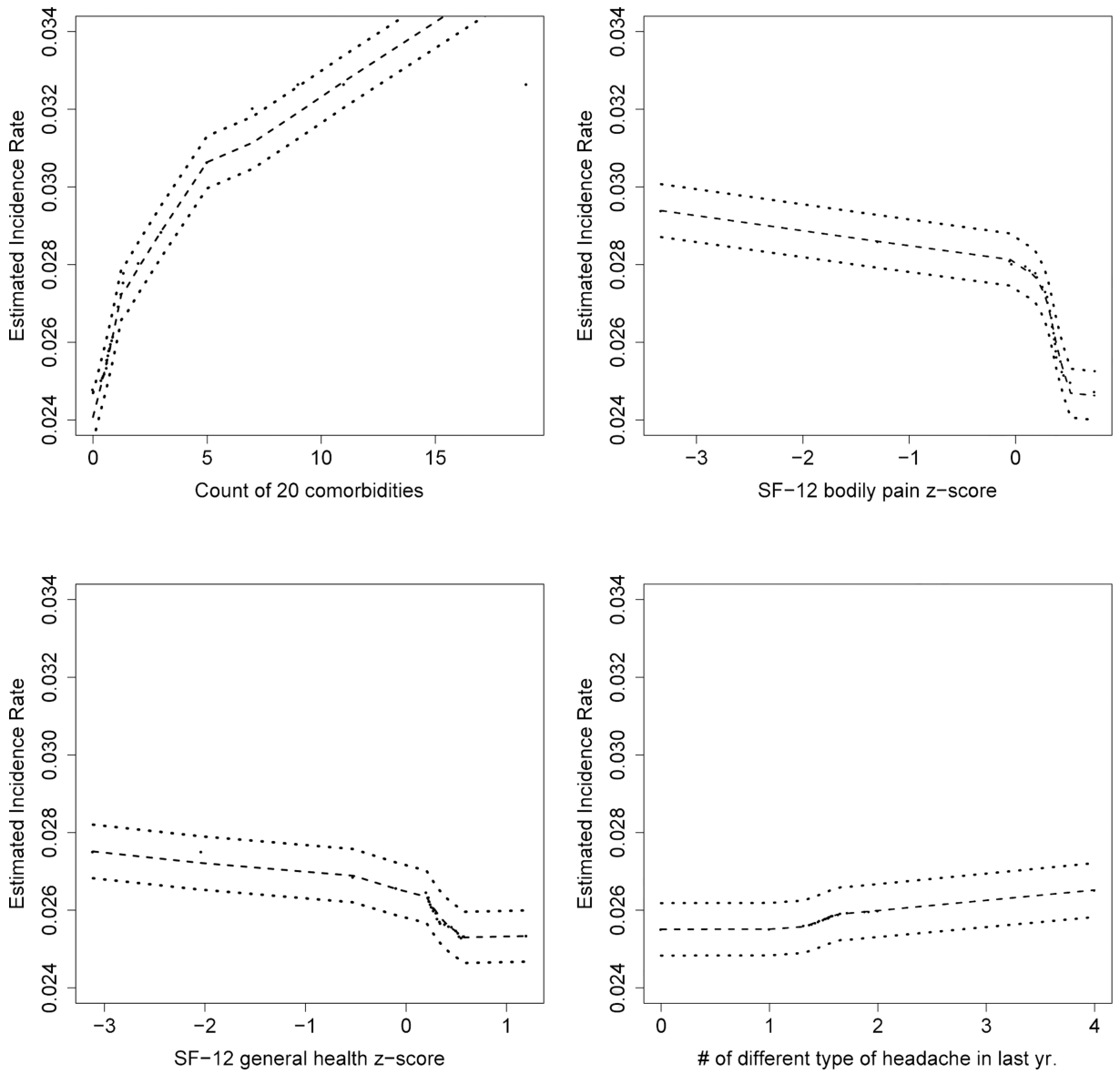


Figure 2. Partial dependence plots for selected health status variables, which show the estimated TMD incidence rate for several possible values of each variable after adjusting for all other OPPERA variables.

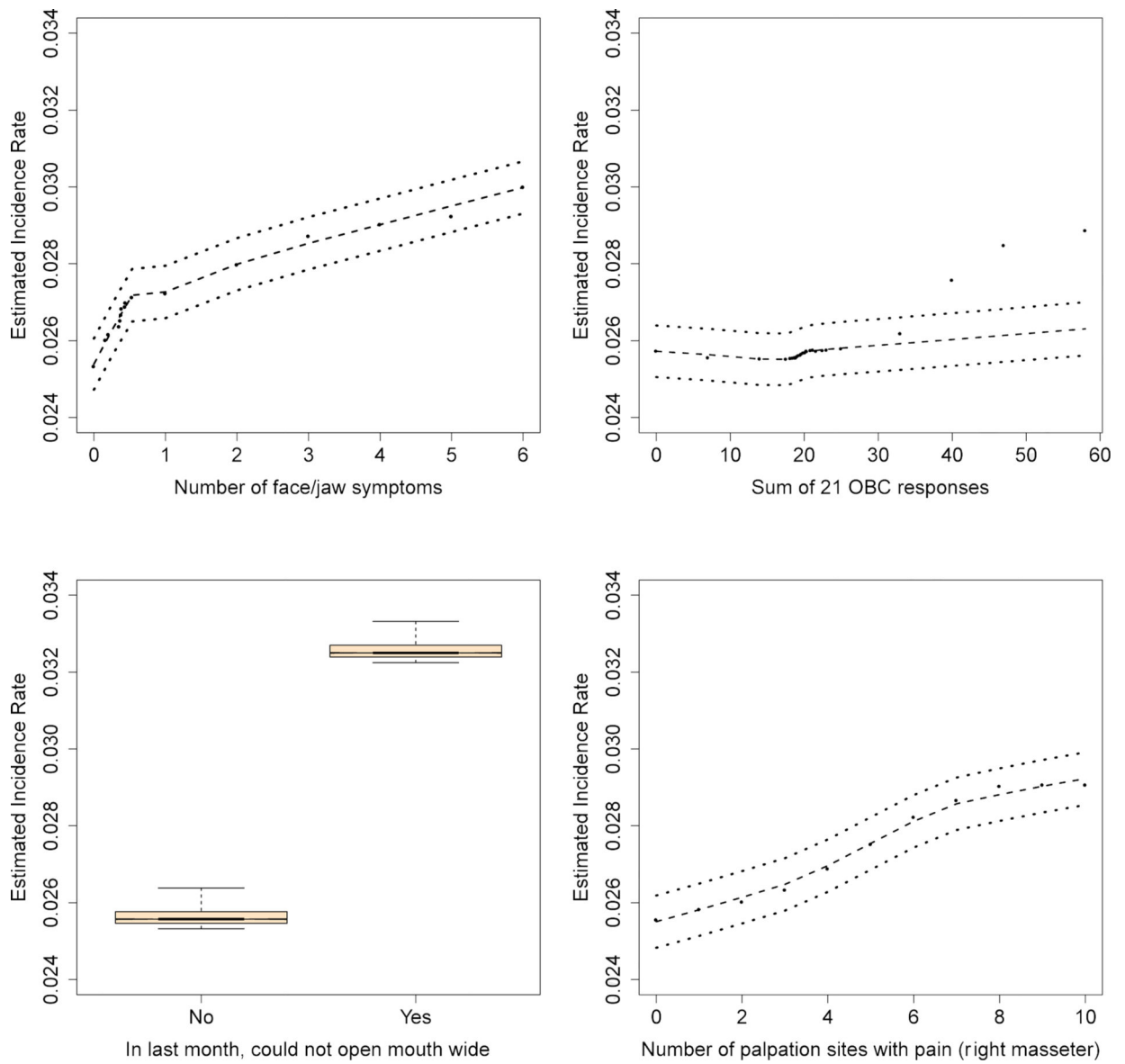


Figure 3. Partial dependence plots for selected clinical variables, which show the estimated TMD incidence rate for several possible values of each variable after adjusting for all other OPPERA variables.

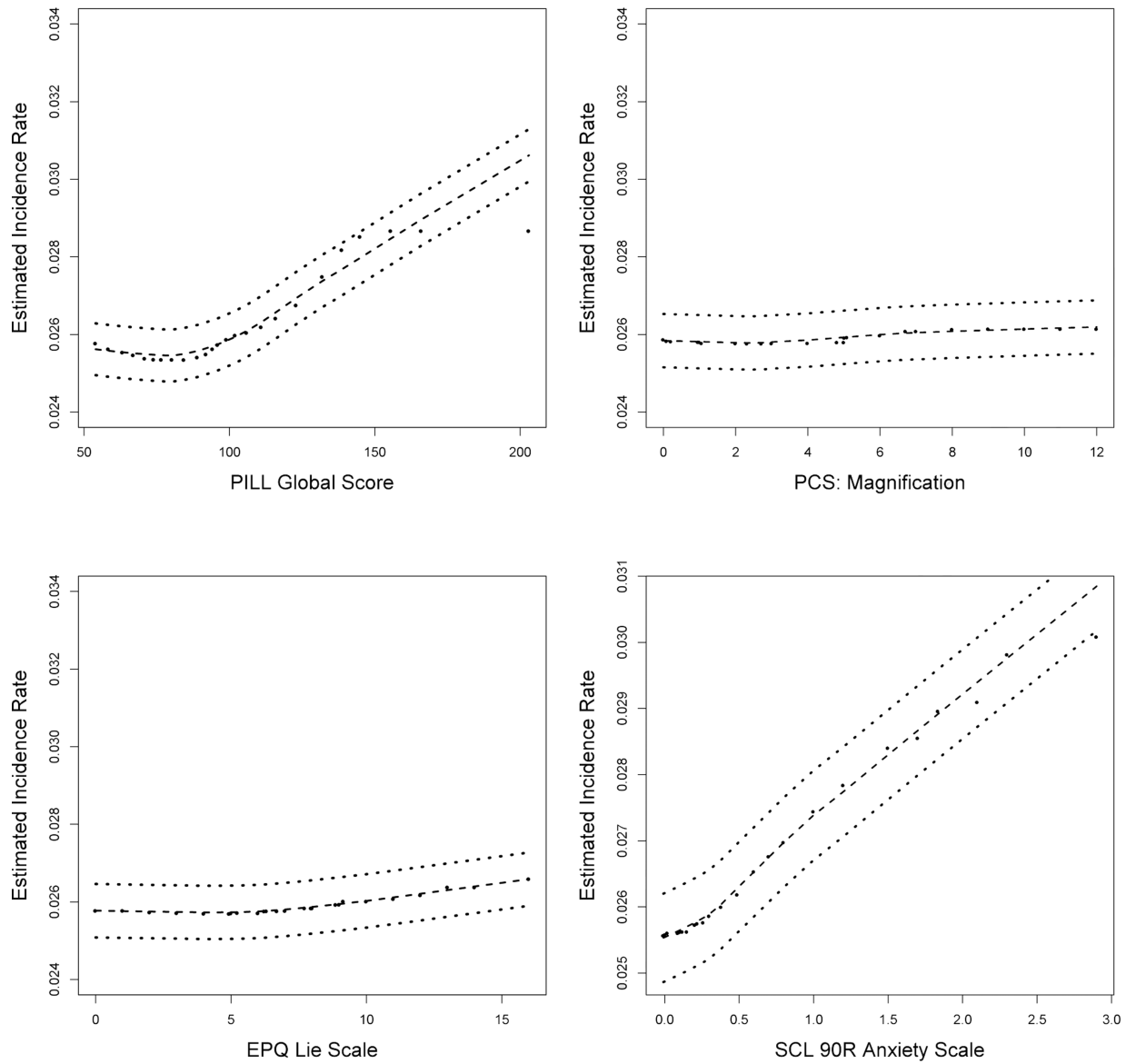


Figure 4. Partial dependence plots for selected psychosocial variables, which show the estimated TMD incidence rate for several possible values of each variable after adjusting for all other OPPERA variables. See Supplementary Figure 7 for a version of this figure with the y-axes redrawn to show additional detail.

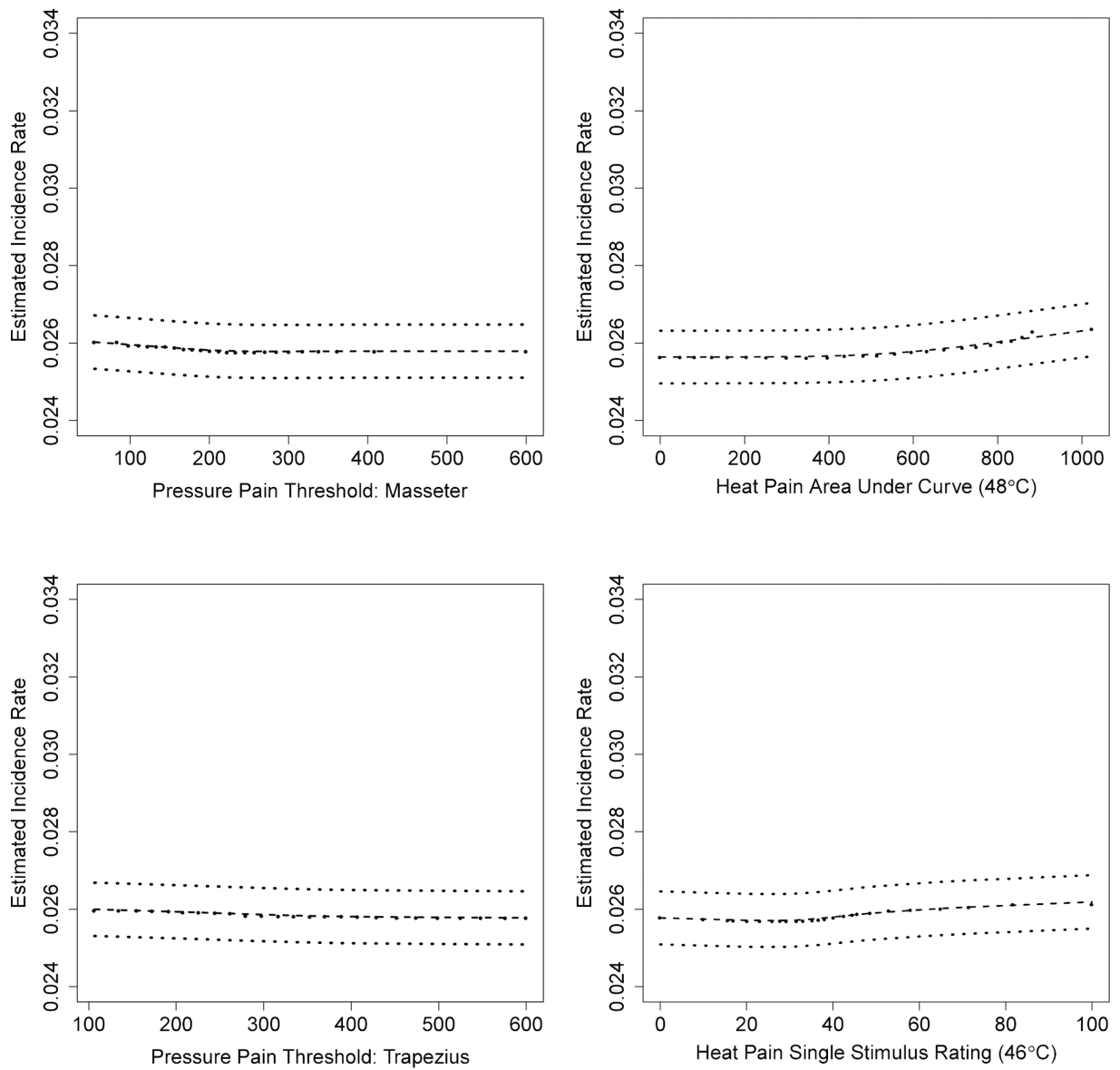


Figure 5. Partial dependence plots for selected pain sensitivity variables, which show the estimated TMD incidence rate for several possible values of each variable after adjusting for all other OPPERA variables. See Supplementary Figure 8 for a version of this figure with the y-axes redrawn to show additional detail.

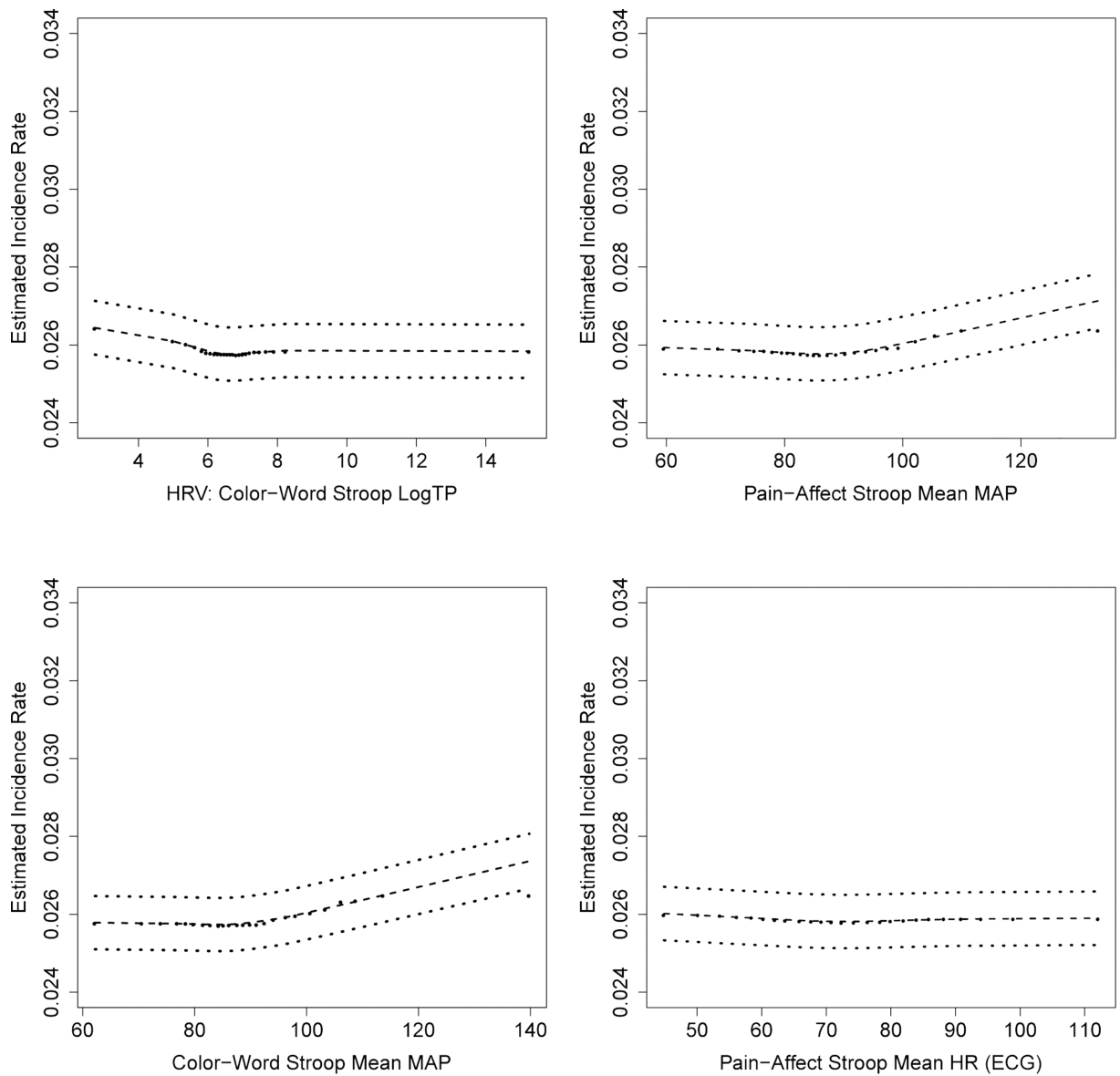


Figure 6. Partial dependence plots for selected autonomic variables, which show the estimated TMD incidence rate for several possible values of each variable after adjusting for all other OPPERA variables. See Supplementary Figure 10 for a version of this figure with the y-axes redrawn to show additional detail.

Table 1

Lasso Regression Coefficients

Variable	Standardized Hazard Ratio	Hazard Ratio
somatization (SCL 90R)	1.180	
count of 20 comorbid conditions	1.123	
count of 6 non-specific orofacial symptoms	1.071	
global sleep score (PSQI)	1.069	
study site (Florida)	1.035	1.057
# of palpation sites with pain (right temporalis)	1.029	
bodily pain (SF-12v2)	0.975	
smoking history (never)	0.984	0.991
lifetime U.S. Residence (less than all my life)	0.985	0.967
average diastolic BP (orthostatic challenge)	1.013	
# of painful anatomical locations during protrusion	1.009	
# of palpation sites with pain (left temporalis)	1.008	
negative impact of life events (LES)	1.005	
general health (SF-12v2)	0.996	
count of 10 IBS symptoms	1.003	
# of palpation sites with pain (left TM joint)	1.003	

Table 2

Concordance Indices for the Random Forest Models Fit to Each OPPERA Domain

Domain	Concordance Index
Cross-Domain	0.74
Autonomic	0.65
Clinical	0.72
Demographic	0.69
Health Status	0.75
Psychosocial	0.73
QST	0.67

Table 3

Putative TMD Risk Factors with the Largest Importance Scores

Variable	Importance Score
Count of 20 comorbid conditions	100.0
Count of 6 non-specific orofacial symptoms	92.9
Study site	90.7
Bodily pain (SF-12v2)	80.6
Oral parafunction sum score (OBC)	66.0
Could not open mouth wide in the last month	54.1
Age	51.6
# of palpation sites with pain (right masseter)	50.0
Marital status	44.7
Somatic symptom reporting (PILL)	42.4
General health (SF-12v2)	31.8
Ever had orthodontic procedures	29.3
Race	25.1
# of palpation sites with pain (left masseter)	23.0
HRV: total power (color-word Stroop)	19.1
# of painful anatomical locations during protrusion	16.3
Average mean arterial pressure (pain-affect Stroop)	16.2
# of different types of headaches in the last year	16.1
Average mean arterial pressure (color-word Stroop)	15.8
Pain with TMJ noises in the past month	15.4
Sleep latency (PSQI)	12.7
Average heart rate - ECG (pain-affect Stroop)	12.6
Lifetime U.S. residence	12.4
Count of 10 IBS symptoms	11.9
Functional limitation in jaw opening (JFLS)	11.5
Self-rated general health	11.1
Could not open mouth wide prior to 1 month ago	10.8
HRV: total power (pain-affect Stroop)	10.8
# of painful anatomical locations during right lateral excursion	10.6
Catastrophizing - magnification (PCS)	10.4

Table 4

Putative TMD Risk Factors with the Largest Importance Scores (by Domain)

Domain	Variable	Importance Score
Autonomic	HRV: total power (color-word Stroop)	19.1
	Average mean arterial pressure (pain-affect Stroop)	16.2
	Average mean arterial pressure (color-word Stroop)	15.8
	Average heart rate - ECG (pain-affect Stroop)	12.6
	HRV: total power (pain-affect Stroop)	10.8
Clinical	Count of non-specific orofacial symptoms	92.9
	Oral parafunction sum score (OBC)	66.0
	Could not open mouth wide in the last month	54.1
	# of palpation sites with pain (right masseter)	50.0
	Ever had orthodontic procedures	29.3
Demographic	Age	51.6
	Marital status	44.7
	Race	25.1
	Lifetime U.S. residence	12.4
	Satisfaction with financial situation	5.5
Health Status	Count of 20 comorbid conditions	100.0
	Bodily pain (SF-12v2)	80.6
	General health (SF-12v2)	31.8
	# of different types of headaches in the last year	16.1
	Sleep latency (PSQI)	12.7
Pain Sensitivity	Pressure pain threshold (masseter)	5.8
	Heat pain ratings of 10 stimuli: area under curve (48°C)	4.2
	Pressure pain threshold (trapezius)	3.7
	Thermal pain single stimulus rating (46°C)	3.6
	Thermal pain single stimulus rating (48°C)	3.5
Psychosocial	Somatic symptom reporting (PILL)	42.4
	Catastrophizing - magnification (PCS)	10.4
	EPQ Lie scale	9.9
	Anxiety (SCL 90R)	9.7
	Mood - clearheaded/confused (POMS-Bi)	6.5