American Pain Society
RESEARCH
EDUCATION
TREATMENT
ADVOCACY
PUBLISHED BY ELSEVIER
The Journal of Pain, Vol 14, No 12 (December), Suppl. 2, 2013: pp T2-T19
Available online at www.jpain.org and www.sciencedirect.com

Study Protocol, Sample Characteristics, and Loss to Follow-Up: The OPPERA Prospective Cohort Study

Eric Bair,[*,†,‡] Naomi C. Brownstein,[‡] Richard Ohrbach,[§] Joel D. Greenspan,[||,¶,**] Ronald Dubner,[||,¶,**] Roger B. Fillingim,[††] William Maixner,[*,†,‡‡] Shad B. Smith,[*,†] Luda Diatchenko,[*,†,§§,||||] Yoly Gonzalez,[§] Sharon M. Gordon,[||,**] Pei-Feng Lim,[*,†] Margarete Ribeiro-Dasilva,[¶¶] Dawn Dampier,[***] Charles Knott,[***] and Gary D. Slade[*,†††,‡‡‡]

*Regional Center for Neurosensory Disorders, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina. Departments of †Endodontics, ‡Biostatistics, ‡‡Pharmacology, †††Dental Ecology, and ‡‡‡Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.
§Department of Oral Diagnostic Sciences, University at Buffalo, Buffalo, New York.
Departments of ||Oral and Maxillofacial Surgery and ¶Neural and Pain Sciences, and **Brotman Facial Pain Center, University of Maryland School of Dentistry, Baltimore, Maryland.
††Department of Community Dentistry and Behavioral Science, University of Florida, College of Dentistry, and Pain Research and Intervention Center of Excellence, Gainesville, Florida.
§§Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.
||||Department of Anesthesia, and Alan Edwards Centre for Resarch on Pain, McGill University, Montreal, Quebec, Canada.
¶¶Department of Restorative Dental Sciences–Divison of Prosthodontics, University of Florida, College of Dentistry, Gainesville, Florida.
***Battelle Memorial Institute, Durham, North Carolina.

Abstract: **When studying incidence of pain conditions such as temporomandibular disorder (TMD), repeated monitoring is needed in prospective cohort studies. However, monitoring methods usually have limitations and, over a period of years, some loss to follow-up is inevitable. The OPPERA prospective cohort study of first-onset TMD screened for symptoms using quarterly questionnaires and examined symptomatic participants to definitively ascertain TMD incidence. During the median 2.8-year observation period, 16% of the 3,263 enrollees completed no follow-up questionnaires, others provided incomplete follow-up, and examinations were not conducted for one third of symptomatic episodes. Although screening methods and examinations were found to have excellent reliability and validity, they were not perfect. Loss to follow-up varied according to some putative TMD risk factors, although multiple imputation to correct the problem suggested that bias was minimal. A second method of multiple imputation that evaluated bias associated with omitted and dubious examinations revealed a slight underestimate of incidence and some small biases in hazard ratios used to quantify effects of risk factors. Although "bottom line" statistical conclusions were not affected, multiply-imputed estimates should be considered when evaluating the large number of risk factors under investigation in the OPPERA study.**

Perspective: ***These findings support the validity of the OPPERA prospective cohort study for the purpose of investigating the etiology of first-onset TMD, providing the foundation for other papers investigating risk factors hypothesized in the OPPERA project.***

Key words: *Temporomandibular disorder, cohort studies, population statistics, epidemiologic methods, proportional hazards models.*

Publication of this supplement was made possible with support of the National Institutes of Health grant U01DE017018 and T32ES007018. The OPPERA program also acknowledges resources specifically provided for this project by the participating institutions: Battelle Memorial Institute; University at Buffalo; University of Florida; University of Maryland; University of North Carolina at Chapel Hill.

R.B.F. and G.D.S. are consultants and equity stock holders, and W.M. and L.D. are cofounders and equity stock holders in Algynomics, Inc, a company providing research services in personalized pain medication and diagnostics. Other authors declare no conflicts of interest.

Supplementary data accompanying this article are available online at www.jpain.org and www.sciencedirect.com.

Address reprint requests to Gary D. Slade, BDSc, DDPH, PhD, Room 4501E, Koury Oral Health Sciences, UNC School of Dentistry, 385 South Columbia Street, CB#7455, Chapel Hill, NC 27599-7455. E-mail: gary_slade@dentistry.unc.edu

1526-5900/$36.00

http://dx.doi.org/10.1016/j.jpain.2013.06.006

The high quality of evidence from prospective cohort studies is attributable to 2 features of the study design. First, by quantifying the likelihood of developing an illness, prospective cohort studies address a critical question asked by individuals who have yet to develop it: "Am I likely to get the illness?" At a population level, the same prognostic information is essential when predicting the impact of interventions on the public health. Second, because exposure to hypothesized risk factors is measured prior to illness onset, prospective cohort studies establish a temporal sequence between putative cause and effect. This is a cardinal criterion for causal inference.[18]

Despite these attributes, prospective cohort studies are relatively uncommon compared to other study designs investigating etiology, primarily because longitudinal data collection is logistically complex, time-consuming, and expensive. Prospective cohort studies of chronic pain face additional challenges[5] because symptom episodes often are transient and recurrent.[3,8,9,27,36-38] Furthermore, because long-term recall of such episodes is unreliable,[17] prospective cohort studies are likely to underestimate incidence of chronic pain when it is recalled only at a single follow-up assessment conducted years after enrollment.

Problems created by poor recall of pain episodes can be addressed by repeated monitoring throughout the study's period of follow-up. The strategy is epitomized in clinical trials where daily diaries monitor pain intensity and symptoms. Daily monitoring is not feasible for studies with lengthy follow-up, and neither is it necessary in population-based studies where the focus is on people's experience of discrete episodes of pain. Instead, monthly or quarterly (3-month) monitoring is appropriate. For example, in a UK-based study,[8] questionnaires were administered monthly for up to 6 months after enrollment of 342 adult patients who had low back pain. During follow-up, 43% reported pain that varied between months. In a U.S. population-based study of 1,336 adolescents,[9] questionnaires administered once every 3 months for up to 3 years evaluated back pain, headache, stomach pain, and facial pain. Forty percent developed 1 or more types of pain, although only 12% had persistent pain. In both studies, intermittent episodes of pain would have been overlooked if only a single follow-up assessment had been conducted at the end of the study.

Repeated monitoring has associated problems of its own, including loss to follow-up, which is almost inevitable in population-based studies. When follow-up assessments are missing, biased estimates are likely if the data are analyzed using conventional methods of complete-case analysis.[22] If loss to follow-up is unrelated to putative risk factors, estimated associations between risk factors and incidence are biased toward the null. A more serious problem occurs when loss to follow-up varies according to putative risk factors. Under such conditions, estimates of association become biased in directions that cannot readily be determined.

This paper reports methodological details of a prospective cohort study of first-onset temporomandibular disorder (TMD), a painful musculoskeletal condition of the face and jaws. The study was part of the project entitled Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA). One aim of this paper is to describe the study's methods of enrollment, follow-up, and ascertainment of TMD incidence. The other aim is to evaluate the degree to which loss to follow-up and related problems in data collection might have affected estimates of the study's 2 main outcome measures: TMD incidence and hazard ratios of association between putative risk factors and TMD incidence.

Methods

The OPPERA prospective cohort study was designed to investigate the etiology of first-onset, painful TMD. Relationships among components of the OPPERA project are depicted in Supplementary e-Fig 1. Seven other papers in this issue present associations between putative risk domains and TMD incidence. Previous publications[24] reported findings from the OPPERA baseline case-control study of chronic TMD. Methods of baseline data collection for this prospective cohort study were provided in one of those papers[32] and are summarized below to provide a background to methods used for follow-up and measurement of TMD incidence.

Institutional review boards at each study site approved study procedures, and participants provided signed, informed consent. The OPPERA study is being conducted under the auspices of a Certificate of Confidentiality (NIDCR-06-17) between the National Institutes of Health and Dr. William Maixner, Program Director of OPPERA. The Certificate protects the privacy of research participants.

Target Population and Selection of Study Participants

The target population was community-based volunteers aged 18 to 44 years with no significant history of TMD who lived or worked near 4 recruitment sites that were selected to provide a demographically diverse sample of U.S. adults: Baltimore, MD; Buffalo, NY; Chapel Hill, NC; and Gainesville, FL. Between May 2006 and November 2008, potential study participants were recruited using advertisements, e-mails, and flyers. The goal was to enroll an inception cohort, defined as people with no current or previous experience of TMD. An initial set of inclusion and exclusion criteria were administered by telephone screening interviews. Inclusion criteria were all 7 of 1) age 18 to 44 years, 2) planning on living in the area for the next 2 years, 3) fluent in written and spoken English, 4) reported having never been diagnosed with TMD, 5) reporting no significant history of orofacial pain (ie, no orofacial pain in the month before enrollment and, prior to that period, no more than 4 days of orofacial pain per month), 6) reporting <5 headaches/month in the 3 months before enrollment, and 7) no reported use of a night guard or occlusal splint. Exclusion criteria were any 1 of 13 conditions: 1) traumatic

facial injury or surgery on the face or jaw within the 6 months preceding enrollment, 2) currently receiving orthodontic treatment, 3) pregnant or nursing, 4) kidney failure or renal dialysis, 5) heart disease or heart failure, 6) chronic respiratory disease that is not controlled with medication, 7) hypertension that is not controlled with medication, 8) epilepsy or medication to control grand mal seizures, 9) hyperthyroidism, 10) diabetes that is not controlled with medication or diet, 11) drug or alcohol abuse, 12) psychiatric disorders or conditions that have required hospitalization, or 13) chemotherapy or radiation therapy.

Our justification for selecting 18- to 44-year-olds was that we expected them to have a higher incidence of TMD than other age groups[7] and therefore represent a high-priority group from a public health perspective. Nearby residency was an important requirement for efficient conduct of the study, and English-language proficiency was necessary because many of the standardized questionnaires were not available in other languages. A negative history of significant TMD symptoms or treatment was necessary to create an inception cohort in which development of first-onset TMD could be monitored. Given the potential overlap of TMD and headache symptoms, we likewise excluded people with frequently occurring headache. Current orthodontic treatment and recent facial injury or surgery were exclusion criteria because we felt that thorough baseline oral examinations would be compromised in such people. Likewise, people with the other health-related conditions were excluded either because the investigators felt the conditions might invalidate components of the baseline clinical assessments or because standard operating procedures of the dental schools at each study site precluded elective dental care for such people.

After screening, potential participants attended a study site's research clinic where 2 additional inclusion criteria were determined: 1) pain reported in the examiner-defined orofacial region for no more than 4 days in the prior 30 days and 2) absence of both TMD myalgia and arthralgia. The latter 2 criteria were determined by trained and calibrated examiners using a protocol based on the Research Diagnostic Criteria for TMD[10] and described in detail elsewhere.[29] In summary, the study participant's orofacial region was defined as examiners touched the following anatomic areas bilaterally: temporalis, preauricular, masseter, posterior mandibular, and submandibular areas. Examiners asked structured questions about any pain history in this defined orofacial region. They also evaluated signs of painful TMD during jaw movement and digital palpation of study participants' orofacial structures. The origin of any pain reported during jaw movement and palpation was classified by the examiner into 1 or more of 5 anatomic locations, each considered bilaterally: temporalis, masseter, lateral pterygoid, submandibular and posterior mandibular, and temporomandibular joint. TMD was classified for study participants who reported both 1) pain in the orofacial region for ≥5 days in the prior 30 days and 2) pain in ≥3 muscles locations (myalgia) or in ≥1 TM joints (arthralgia) during jaw movement or orofacial palpation. Although participants with examiner-verified TMD were excluded from the inception cohort, individuals were retained in the inception cohort if they reported pain during examination procedures but did not report a history of pain for ≥5 days in the prior 30 days.

Baseline Interview, Questionnaires, Examination, Physiologic Testing, and Biospecimen Collection

Prior to the clinic visit, study participants completed psychosocial questionnaires—either paper forms sent by postal mail or equivalent versions online.[12] During the clinic visit, additional psychosocial and health status questionnaires were completed. In addition to the clinical orofacial assessments of TMD, examiners evaluated study participants' tenderness to neck and body palpation.[29] Quantitative sensory testing performed after the examination measured responses to standardized noxious stimuli.[16] Autonomic function was monitored at rest, during orthostatic challenge, and during the Stroop color-word test and pain-affect test.[25] Anthropometric measurements were recorded, and a 20-mL sample of peripheral blood was collected by venipuncture for subsequent DNA purification and genotyping.[35]

Examiner Training, Calibration, and Reliability

Prior to study initiation, clinical examiners from each study site were trained and jointly calibrated by 2 expert dentists (Y.G. and R.O.). Y.G. served as the reference examiner throughout the study, and R.O. monitored examiner performance for adherence to protocol. In a separate session, examiners conducted at least 10 blinded, replicated examinations of non-OPPERA volunteers: 1 examination in each pair was conducted by the OPPERA examiner and the other by the reference examiner. Volunteers in these reliability studies were selected to yield an approximate 2:1 ratio of TMD cases and noncases. Similar calibration and reliability sessions were repeated approximately annually after study initiation. Data from the blinded, replicate examinations were analyzed for interexaminer reliability computed using the Kappa statistic.

Follow-Up Quarterly Health Update Questionnaires

At 3-month intervals after enrollment, study participants completed a quarterly health update questionnaire that included screening questions about orofacial pain. In order to capture an approximate 3-month period of symptoms, the questionnaire was imprinted with an individualized reference date that was 13 weeks prior to the intended completion date. Two weeks prior to the intended completion date, the questionnaire was sent to the study participant with instruction to answer the questions on paper or online. Reminder e-mails were sent 1 week before and on the day of the intended completion date.

Screening questions asked about pain symptoms and other health-related events in the period since the reference date. The intended reporting period of 13 weeks varied in duration because some respondents completed the quarterly health update earlier or later than intended. Questions that screened for likely onset of TMD inquired about "headaches or pain in your face, jaw, temples, in front of the ear, or in the ear" (hereafter "orofacial pain") during the reporting period. An orofacial symptom episode was defined when answers to questions about density of orofacial pain symptoms met either of 2 criteria: 1) ≥5 consecutive days of orofacial pain per month for ≥2 months, with ≥1 day of orofacial pain in the 2 weeks preceding questionnaire completion, or 2) ≥5 consecutive days of orofacial pain in the month preceding questionnaire completion, with ≥5 days of orofacial pain in the 2 weeks preceding questionnaire completion. Reporting periods that did not meet this threshold are referred to hereafter as "asymptomatic episodes."

If a study participant did not complete any quarterly health updates within a 12-month period or if his or her contact information became invalid, he or she was traced using an iterative process. First, local site coordinators used available contact information, including alternative point-of-contact information collected at enrollment, to locate study participants. If this effort yielded no new contact information, the data coordination center traced participants using a variety of professional epidemiologic tracing resources.

Study participants were paid a $5 incentive for each completed quarterly health update, with an additional $10 bonus after completion of every fourth consecutive questionnaire. Consistent completion of questionnaires over a 5-year period earned the participants an additional $50 bonus.

Follow-Up Clinical Assessments and Classification of First-Onset TMD

Study participants reporting an orofacial symptom episode were asked to attend the research clinic where examiners used the same baseline examination protocol to classify presence or absence of TMD. In this way, participants were classified with first-onset TMD when they met each of 2 criteria: 1) ≥5 days of pain during the preceding 30 days in TMD locations specified by examiner and 2) examiner findings of pain in ≥1 temporomandibular joint(s) (arthralgia) or in ≥3 muscle locations (myalgia) during jaw maneuver or palpation. Most of the other measures recorded at baseline were repeated during these follow-up visits of symptomatic study participants (Supplementary e-Table 1).

Follow-up clinical assessments were also conducted for a random sample of study participants who had asymptomatic episodes (Supplementary e-Fig 1). It was a matched sampling design that selected 1 asymptomatic participant at random to correspond with each symptomatic participant. There were 4 matching criteria: 1) enrollment within 15 days of one another, 2) enrollment at the same study site, 3) same gender, and 4) quarterly questionnaires completed within 3 months of one another. One goal of examining asymptomatic participants was to evaluate the quarterly health update's negative predictive value, defined as the percentage of asymptomatic episodes that were classified as noncases of TMD during the subsequent examination. Most of the other measures recorded at baseline were repeated during these clinic visits that followed asymptomatic episodes (Supplementary e-Table 1).

Most study participants who made follow-up visits attended the research clinic at the study site where they were enrolled, although 6 participants were reexamined at a different study site after relocating. In results reported here and elsewhere, the participant's study site was deemed to be the site at which he or she was enrolled.

Reliability of Follow-Up Screening Questionnaires

Consistency of responses to quarterly health updates was evaluated in a separate study of test-retest reliability conducted at each of the study sites. Approximately equal numbers of TMD cases and TMD-free controls were recruited. A total of 105 participants completed 2 quarterly health updates separated by an interval of 4 to 8 days. Kappa statistics were computed as measures of test-retest reliability in classifying orofacial symptom episodes.

Quality Control and Quality Assessment of Classification of First-Onset TMD

Each follow-up examination was subjected to 3 steps in data quality control.[32] All examination forms were reviewed by an OPPERA expert dentist at each site to verify consistency between findings and case classification and to evaluate any notes recorded by the study examiner that might influence case classification. When required, a clinical review panel was conducted by conference call; the panel was composed of expert dentists and the OPPERA principal investigator expert in TMD case classification (R.O.). They reviewed online copies of casebook forms and resolved instances of uncertain case classification. Finally, a software algorithm used data from scanned examination forms to verify that TMD case classification was consistent with findings for relevant muscles and joints recorded during jaw movement and palpation procedures. Any discrepancies between the algorithmically derived classification and the examiner's classification were reviewed and resolved by members of the expert pain panel.

Quality assessment of examination findings was undertaken annually by merging data from quarterly health updates and examinations to compute negative and positive predictive values. The latter was defined as the percentage of symptomatic episodes that were classified as first-onset TMD during the subsequent examination. The findings were reported to the OPPERA External Scientific Advisory Committee that monitored the study's progress.

Sample Size Determination for the Inception Cohort

OPPERA was designed with a target sample size of 3,200 enrolled study participants. This was expected to yield 196 cases of first-onset TMD during a 3-year follow-up period, assuming 30% loss to follow-up. The expectations were based on incidence and cohort retention rates observed in a previous study conducted at the North Carolina study site.[6] This target sample size was calculated to provide 80% statistical power to detect risk ratios of at least 1.8 for risk predictors with as few as 15% of people in the high-risk category. This was consistent with the magnitude of effect seen for genetic predictors in the previous North Carolina study.[6]

Statistical Methods

The follow-up period for each study participant was computed as the time from the enrollment date to the first of 3 possible events: 1) date of the examination when first-onset TMD was classified, 2) date of the last-completed quarterly health update for people who stopped returning quarterly health updates before May 31, 2011, or 3) the census date used for this analysis (ie, May 31, 2011). The first event represented an observed outcome whereas the other 2 events represent censoring. If a participant had gaps in follow-up where he or she failed to complete a quarterly questionnaire, the period was nonetheless included in the participant's total period of follow-up, so long as a subsequent questionnaire was completed prior to the census date. Additionally, it was assumed that the participant did not develop TMD during any such gaps. Although the 260 incident cases continued to complete quarterly health updates after developing TMD (Supplementary e-Fig 1), those observation periods were not used in this analysis. Participants who were examined but found not to have TMD likewise continued to complete quarterly health updates, and those observation periods did contribute to this analysis.

Using these criteria, the average annual rate of first-onset TMD was calculated as the number of people with first-onset TMD divided by the sum of follow-up periods. The result was expressed as the percentage of people per annum (equivalent to the number of incident cases per 100 person-years of follow-up). For descriptive purposes, an adjusted average annual incidence was computed using a Poisson regression model that adjusted for study site using the Buffalo study site as the referent.

To test hypotheses about associations between baseline risk factors and TMD incidence, hazard ratios were computed using Cox proportional hazard regression models. Hazard ratios, the relative difference in hazard rates between 2 groups, are a theoretical construct because the hazard rate is an unobservable, instantaneous event rate as the duration of follow-up approaches zero. However, hazard ratios are a good approximation of the incidence rate ratio in a cohort study.[31] Although incidence rate ratios can be modeled statistically, Cox models require fewer statistical assumptions and therefore were adopted as the standard method to test for associations with putative risk factors. Hereafter, we use the term *incidence* when referring both to the annual incidence rate and the hazard rate.

For the Cox models, an incident case was regarded as an event; otherwise participants were censored. Each participant's follow-up period (defined above) was used as the time-to-event. When the baseline risk factor was categorical, one category was nominated as the referent and indicator variables represented each of the other categories. The requirement of proportional hazards was evaluated for each putative risk factor by testing the null hypothesis of no correlation between the scaled Schoenfeld residuals of the appropriate coefficient and (Kaplan-Meier transformed) time.[14] Quantile-quantile plots of the resulting *P* values were generated and the false-discovery rate was computed to identify any characteristics that departed markedly from the assumption of proportional hazards (Fig 1).

The conventional regression models described above can produce erroneous results when there are certain patterns of missing data for variables used in the analysis. Specifically, if the probability of having a missing value depends on the unobserved value, the data are said to be "missing not at random."[22] Because conventional regression methods use only the observed data, ignoring the pattern of missing data, the estimates therefore are generally biased when the data are missing not at random. Only under carefully considered assumptions can one obtain unbiased regression estimates without further corrections. Two such conditions are the following: 1) the probability of having a missing value is independent of the data (ie, "missing completely at random") or 2) the probability of having a missing value is independent of the missing values but may depend on the observed values or on observed covariates (ie, "missing at random"). In

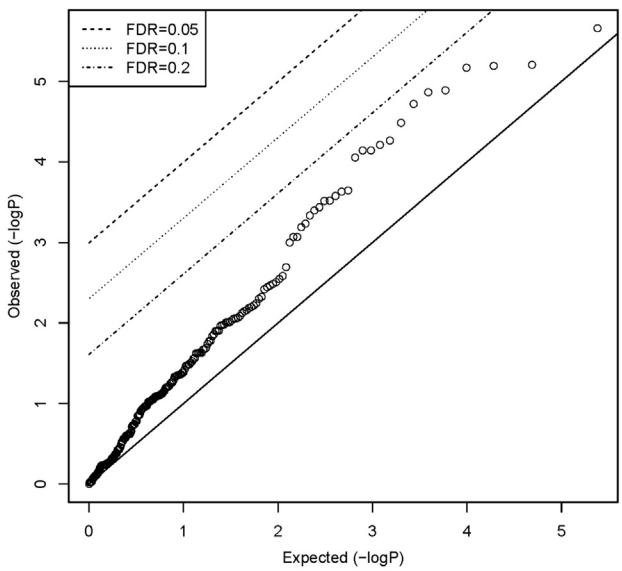

Figure 1. Tests of proportional hazards assumption: OPPERA prospective cohort study, 2006 to 2011. Quantile-quantile plots of *P* values from 251 tests of proportional hazards assumption. FDR, false-discovery rate.

practice, it is virtually impossible to determine if the data truly are missing at random, so it is prudent to conduct sensitivity analysis evaluating the impact of missing data.

Four types of sensitivity analysis were undertaken to evaluate potential bias associated with incomplete follow-up. Baseline characteristics were first compared between participants who completed one or more quarterly health updates and participants who completed no quarterly health updates (the latter are defined here as participants with complete loss to follow-up). Differences were evaluated using Student's t-test for continuous measures and likelihood ratio chi-square tests for categorical measures. Quantile-quantile plots of the resulting *P* values were generated for 4 risk factor domains hypothesized in the OPPERA heuristic model[24]: psychosocial characteristics; quantitative measures of pain sensitivity; cardiovascular measures of autonomic function; and clinical measures of pain and health status. Within each domain, one characteristic most strongly associated both with cohort retention and with TMD incidence was selected. Those 4 characteristics, together with age, gender, race, and study site, were then used to classify participants in the inception cohort, and 3 measures of cohort retention were compared between subgroups: 1) the percentage of participants with complete follow-up was calculated for each subgroup and results were compared using the chi-square test; 2) the probability of remaining in the cohort for at least 2 years was computed using the life-table method, and results were compared using the likelihood ratio test of the survivor function; and 3) the median number of completed quarterly health updates was computed and results were compared using the Brown-Mood test for median scores. The latter 2 statistics, computed for participants who completed at least 1 quarterly health update, were used as indicators of "partial" loss to follow-up.

The second type of sensitivity analysis used variables identified in the preceding analysis to impute TMD incidence for participants with complete loss to follow-up. Hot-deck multiple imputation[1] was performed, with the goal to create groups composed of people with similar baseline characteristics ("hot decks"), within which observed outcomes from people with follow-up data were used to impute likely outcome values for people with no follow-up data. In principle,[1] the characteristics used to create "similarity" within the hot decks should be characteristics that are associated both with loss to follow-up and with the outcome.

Five steps therefore were used in this method of imputation: 1) Demographic and baseline risk factors identified in the preceding analysis were used in a multivariable binary logistic model regression to predict odds of complete loss to follow-up for all participants in the inception cohort. 2) Participants in the inception cohort were ranked according to the value of the model's linear predictor, from which 20 equal-sized strata were calculated, each with a successively greater probability of complete loss to follow-up. 3) For each of the 521 participants completely lost to follow-up, 1 participant was sampled at random from among all participants in the same stratum who provided follow-up data. The method used simple random sampling with replacement. The sampled participant's follow-up status (first-onset TMD or censored) and period of follow-up were used as the imputed estimates for the individual lost to follow-up. 4) The imputed records were added to the records from 2,737 participants who had follow-up data, creating a data set of 3,258 individuals with complete information about TMD incidence and follow-up period. 5) Steps 3 and 4 were repeated 100 times, with independent random sampling in each replication. Incidence rates and hazard ratios were calculated for each of the 100 replicated data sets. The 100 sets of results were combined using the "mianalyze" procedure in SAS (SAS Institute Inc, Cary, NC) to generate valid estimates of rates, rate ratios, and standard errors.

The third type of sensitivity analysis dealt with 2 problems that occurred in enumerating incident cases: 1) follow-up examinations were not conducted for approximately one third of symptomatic episodes, usually because the participant was unable or unwilling to attend the research clinic for an examination; and 2) case classifications made by one examiner (hereafter, "examiner 4") were deemed dubious. Specifically, as documented below, the examiner's case classifications produced a conspicuously greater positive predictive and a lower negative predictive value than other examiners' findings. The first problem likely contributed to underenumeration of incident cases, whereas the second problem likely contributed to overenumeration. To quantify the net effect, a 2-stage, multiple imputation procedure was developed. In the first stage, a binary logistic generalized linear mixed model regression equation estimated the probability of examiner-verified TMD for all symptomatic episodes that were followed by an examination, excluding examinations conducted by examiner 4. Predictor variables were selected from the participants' baseline characteristics and their responses to the quarterly health update that initiated the examination. The logistic regression parameters were applied to symptomatic episodes that were not followed by an examination and to all episodes that were examined by examiner 4, yielding a predicted probability of first-onset TMD for such episodes. Each episode's predicted probability was then used to generate 100 Bernoulli random variables (1 or 0) signifying an imputed incident TMD case or noncase, respectively. The imputed case classifications were combined with observed follow-up data, and incidence rates and hazard ratios were calculated for each of the 100 replicated data sets. The 100 sets of results were combined using the "mianalyze" procedure in SAS to generate valid estimates of rates, rate ratios, and standard errors. For a more detailed description of the methodology used to impute for missing follow-up examinations, see Brownstein et al.[2]

The fourth type of sensitivity analysis addressed the problem of false negatives from quarterly health updates, defined as asymptomatic episodes that would have been classified as cases of TMD had they been examined. When the proportions are expressed as percentages, the false-negative rate equals 100 minus the negative predictive value. The calculation was made using the observed negative predictive value from follow-up examinations of the random sample of asymptomatic episodes,

described above. The false-negative rate was applied to nonexamined participants who reported asymptomatic episodes, and the expected number of incident cases was added to the observed number of incident cases to provide a single-imputed estimate of TMD incidence.

Results

Enrollment and Follow-Up

Between May 2006 and November 2008, 3,263 participants who did not have TMD were enrolled into the inception cohort (Fig 2). At enrollment, 85% of participants (n = 2,770) reported having never experienced orofacial pain, and the remaining 15% (n = 488) reported some history of orofacial pain that was below the enrollment-exclusion threshold of ≥5 days per month. Postenrollment audits found that 5 were ineligible, and they were excluded from all analysis. Sixteen percent (n = 521) of participants were completely lost to follow-up whereas the remaining 84% (n = 2,737) completed 1 or more quarterly health updates for a total of 26,666 follow-up questionnaires. This represented a median of 10 quarterly health updates per person, somewhat less than the median of 14 questionnaires per person that would have been completed had there been no partial loss to follow-up prior to May 2011. In fact, only 1,154 participants (42% of 2,737) completed all intended quarterly health updates through May 2011. The shortfall between the number of intended questionnaires through May 2011 and the number completed represented the degree of partial loss to follow-up. The median shortfall was 3 questionnaires per person, with lower and upper quartiles of 0 and 9 questionnaires, respectively. The median period of follow-up was 2.8 years per person (minimum = .2 years, maximum = 5.2 years) for a total of 7,403 person years of follow-up.

During follow-up, there were 721 orofacial symptom episodes, of which 478 (66%) were accompanied by examinations that classified 235 participants as cases of first-onset TMD. Of the 25,945 quarterly health updates with asymptomatic episodes, 338 were selected at random for examination, and 25 of the examinations (7%) were classified with first-onset TMD. The interval between quarterly health update and examination varied from 0 to 85 days (median = 14 days) for the 260 cases of first-onset TMD and from 0 to 87 days (median = 23-days) for examinees who did not have TMD. Two thirds (70.4%) of the 260 incident cases reported having experienced TMD symptoms for 1 or 2 months in the 3-month period prior to the examination, and 65% said that their symptoms occurred in recurrent bouts.

Factors Associated With Cohort Retention

The percentage of participants retained in the cohort varied significantly according to gender, race, and study site, although not age (Table 1). Demographic groups and study sites had correspondingly large differences in the probability of retention for 2 years and in the number of completed quarterly health updates. For the full set of 157 baseline predictor variables, at least 5 measures within each of the 4 risk factor domains were associated with complete loss to follow-up to a degree that exceeded chance, as judged by quantile-quantile plots (Fig 3). Summary statistics of all variables are presented in Supplementary e-Table 2 through Supplementary e-Table 5, showing that most ratios of mean values between the group retained in the cohort and the group lost to follow-up varied between .7 and 1.3, whereas most ratios of odds between the 2 groups varied between .5 and 1.7.

Based on these findings and associations with TMD incidence reported in other papers in this issue, the following baseline predictor variables were selected for hot-deck multiple imputation: number of body sites that were tender to palpation (higher values were associated with greater TMD incidence[28]), change in mean arterial blood pressure during the Stroop emotional word task (higher values were associated with lower TMD incidence[15]), pressure pain thresholds measured at the trapezius (higher values were associated with lower TMD incidence[15]), and the Perceived Stress Scale (higher values were associated with greater TMD incidence[11]). Variation in cohort retention across terciles of those variables is summarized in Table 1. In some instances, the factor associated with greater cohort retention (eg, number of tender body sites) was also associated with greater TMD incidence,[28] whereas in other instances, there were opposing effects. For example, greater pressure pain threshold, which was associated with greater cohort retention (Table 1), was associated with lower TMD incidence.[15]

The 4 demographic variables were also selected for hot-deck imputation, based on findings that greater age and African American race were associated with greater TMD incidence, Asians had lower TMD incidence, and females had marginally greater TMD incidence than males.[33] When all 8 of the selected variables were evaluated in a multivariable binary logistic regression model, each made at least a nominal contribution to predicting the probability of complete loss to follow-up (Supplementary e-Table 6). The area under the receiver operating characteristic curve was .69 based on this model, indicating that this set of variables provided a reasonable level of discrimination between people retained in the cohort and participants completely lost to follow-up. To illustrate, there was more than 2-fold variation in TMD incidence rate among the 20 strata, ranked inversely according to predicted probability of retention in the cohort (Supplementary e-Table 7). However, the relationship between probability of retention and TMD incidence was not monotonic because, as noted above, some factors had opposing directions of association with cohort retention and TMD incidence (Supplementary e-Table 7).

Factors Associated With Examination Following TMD Symptom Episodes

From 721 symptom episodes, 478 examinations were completed. The percentage examined did not vary

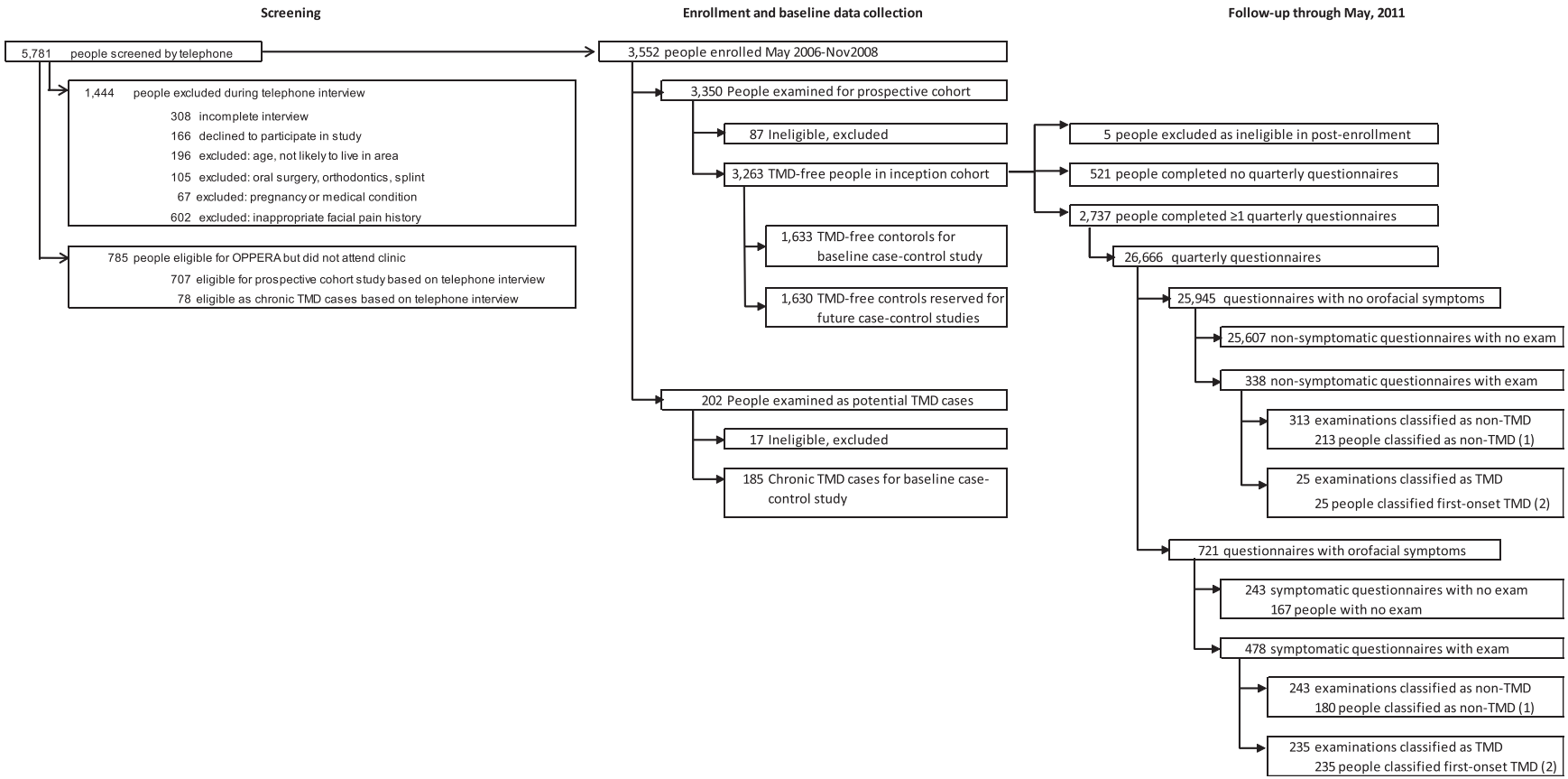

Figure 2. Flowchart of enrollment and follow-up: OPPERA prospective cohort study, 2006 to 2011. *Follow-up data collection continued after examination of people who did not have TMD. Hence, individuals could have more than 1 follow-up examination prior to censoring or developing TMD. †Follow-up data collected after examiner-classified TMD were not used in this analysis.

Table 1. Follow-Up of Participants Through May 2011: OPPERA Prospective Cohort Study, 2006–2011

		People With Follow-Up Data*		Probability of Remaining in Cohort for 2 Years†		Number of Quarterly Health Updates‡	
	Number of People Enrolled	n	%*	Probability (%)	SE	Median	1st, 3rd Quartile
All people	3,258	2,737	84.0	62.1	.9	10	5, 14
Demographics and study site							
Age when enrolled, y							
18–24	1,706	1,421	83.3	62.3	1.2	10	5, 14
25–34	860	736	85.6	64.4	1.6	11	5, 15
35–44	692	580	83.8	58.4	1.9	9.5	4, 14
P value			.320	.007		.003	
Gender							
Female	1,862	1,630	87.5	65.2	1.1	11	5, 15
Male	1,396	1,107	79.3	57.9	1.3	9	4, 14
P value			<.001	<.001		<.001	
Race/ethnicity							
White	1,637	1,448	88.5	69.6	1.1	11	6, 15
Black or African American	1,012	766	75.7	49.6	1.6	7	3, 12
Asian	299	256	85.6	62.5	2.8	10	4, 15
Hispanic	211	178	84.4	62.0	3.3	10	5, 15
Other or unstated	99	89	89.9	60.3	4.9	11	3, 15
P value			<.001	<.001		<.001	
Study site							
Baltimore, MD	768	574	74.7	49.7	1.8	8	3, 12
Buffalo, NY	797	693	87.0	64.9	1.7	10	4, 14
Chapel Hill, NC	815	705	86.5	70.1	1.6	11	6, 15
Gainesville, FL	878	765	87.1	62.5	1.6	11	5, 15
P value			<.001	<.001		<.001	
Risk predictors recorded at baseline							
No. of tender body sites							
None	1,584	1,301	82.1	62.3	1.2	10	5, 14
1–3	808	682	84.4	61.8	1.7	10	5, 15
≥4	866	754	87.1	61.6	1.7	11	4, 15
P value			.006	.920		.150	
Stroop-Emotion Δ MAP,‡ mmHg							
<−2	982	814	82.9	59.3	1.6	10	4, 14
−2 to +2	911	774	85.0	62.5	1.6	10	5, 15
>+2	716	625	87.3	67.4	1.8	11	5, 15
P value			.044	.000		.001	
Pressure pain threshold: trapezius, kPa							
<275	1,082	895	82.7	60.6	1.5	10	5, 14
275–<440	1,089	897	82.4	59.6	1.5	11	4, 15
≥440	1,057	922	87.2	66.4	1.5	10	5, 14
P value			.003	.011		.130	

Table 1. Continued

	Number of People Enrolled	People With Follow-Up Data*		Probability of Remaining in Cohort for 2 Years†		Number of Quarterly Health Updates‡	
		N	%*	Probability (%)	SE	Median	1st, 3rd Quartile
Perceived Stress Scale							
<12	1,035	914	88.3	70.0	1.4	11.0	6, 15
12–<18	1,132	962	85.0	64.5	1.4	10.5	5, 15
≥18	1,064	847	79.6	52.4	1.5	8.0	3, 13
P value			<.001	<.001		<.001	

*P value is from chi-square test of the null hypothesis that percentage with follow-up data is equivalent among subgroups.
†Product-limit survival estimates of probability of retention in cohort. *P* value is from log-rank test of equivalence in probability among subgroups.
‡*P* value is from Brown-Mood test for median scores test of the null hypothesis that the median number of quarterly health updates is equivalent among subgroups.

significantly according to most of the demographic characteristics and baseline risk predictors (Table 2). However, the probability of follow-up examinations varied significantly among study sites, and it was lowest for symptom episodes that occurred some years after enrollment.

Outcomes From Examinations Following Symptomatic and Asymptomatic Quarterly Health Updates

Of the 478 orofacial symptom episodes that were examined, 49.2% were classified as first-onset TMD, and hence the positive predictive value of the screening questions was 49.2% (Table 3). Conversely, 92.6% of the asymptomatic episodes were confirmed as noncases of TMD when examined (hence, the negative predictive value was 92.6%). Positive predictive value varied significantly according to age, race/ethnicity, and study site, whereas the number of tender body sites was the only baseline risk factor associated with positive predictive value. Both nonspecific orofacial pain reported in the quarterly health update and time since enrollment were associated with positive predictive value. In most instances, the same characteristics were associated with variation in negative predictive values.

Of greatest concern were the results from one examiner who registered 100% positive predictive value but only 76.5% negative predictive value (Table 3, Examiner 4). These values differed markedly from the results for other examiners, either at the same study site or elsewhere, and indicated that the examiner was much more likely than other examiners to classify participants as incident cases.

Reliability and Validity of Quarterly Health Updates

In test-retest reliability of the quarterly health update questionnaire, the kappa statistic for reliability of symptom episodes was .83 (95% confidence limits [CLs] = .72, .95), indicating excellent agreement. The quarterly health update's positive predictive value of 49.2% reported above indicated a high rate of false-positive screenings (Table 3, "All people"). However, the screening questions had good overall validity when also considering the high negative predictive value of 92.6%. This corresponded to sensitivity of 90.4% and specificity of 56.3%. After dubious case classifications by examiner 4 were excluded, the screening questions had positive predictive value of 39.7%, negative predictive value of 94.4%, sensitivity of 90.4%, and specificity of 54.1%.

Reliability of Examiners' Classifications of TMD

The 7 study examiners were evaluated for reliability in TMD case classification at multiple sessions during the 5-year duration of the study, yielding a total of 432 paired examinations (Table 4). Kappa statistics ranged from .82 to 1.00, signifying excellent interexaminer

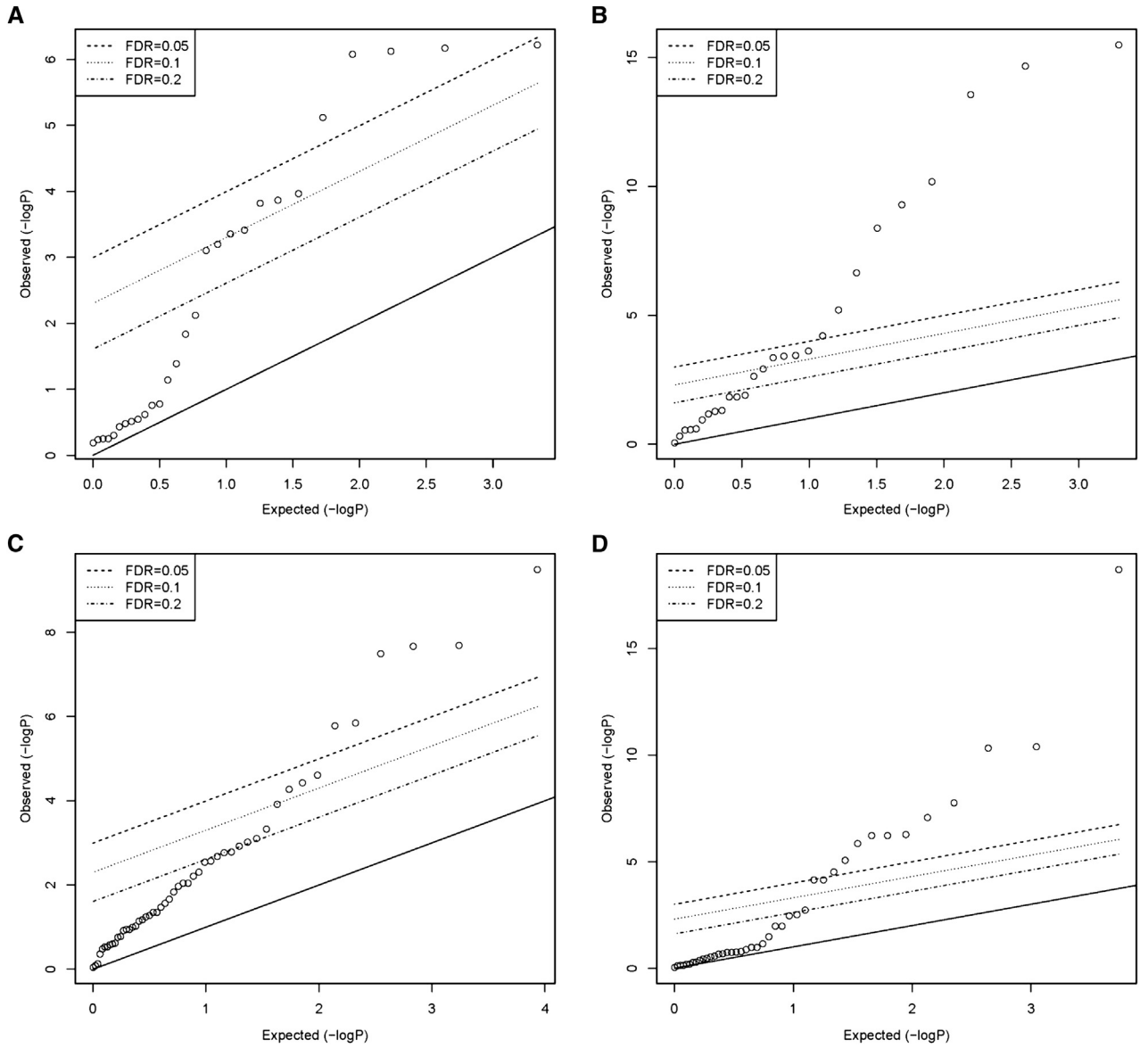

Figure 3. Tests of association between baseline characteristics and loss to follow-up: OPPERA prospective cohort study, 2006 to 2011. Quantile-quantile plots of *P* values from tests of association between 521 people completely lost to follow-up and 2,737 people who completed 1 or more follow-up questionnaires. Dependent variables were 28 measures of pain sensitivity **(A)**, 27 psychological characteristics **(B)**, 51 measures of autonomic function **(C)**, and 42 clinical characteristics **(D)**.

reliability between each of the study site examiners and OPPERA's referent examiner. It was noteworthy that examiner 4 had excellent interexaminer reliability in all 4 reliability assessments through February 2009. (In August 2009, examiner 4 stopped working in the study.)

Sensitivity Analysis Using Complete-Case and Multiply-Imputed Data Sets

In univariate analysis of the complete-case data set of 2,737 participants who completed at least one quarterly health update the TMD annual incidence rate was 3.5%. There were statistically significant differences in incidence according to age, race/ethnicity, study site, and all 4 baseline risk predictors, although not gender (Table 5).

An identical annual incidence rate of 3.5% was obtained using hot-deck multiple imputation, which accounted for complete loss to follow-up (Table 5). The baseline characteristics noted above using complete-case analysis were likewise associated with the imputed incidence rate, whereas gender was not. Compared to the complete case analysis, hazard ratios differed by no more than .2 in absolute value. The largest difference between the analytic methods was observed for the highest tercile of perceived stress, where the imputed hazard ratio of 1.77 (95% CLs = 1.31, 1.40) was lower than the complete-case hazard ratio of 1.96 (95% CLs = 1.44, 2.66). In each instance where hazard ratios differed between the 2 methods, the imputed ratio was closer to the null value of 1.0.

The second method of multiple imputation used parameters from the binary logistic regression model that predicted TMD case classification for symptomatic episodes (Supplementary e-Table 8). When applied to

Table 2. Examinations Following Orofacial Symptom Episodes: OPPERA Prospective Cohort Study, 2006–2011

	Number of Symptom Episodes	Number of Examinations	% Examined
All people	721	478	66.3
Demographics and study site			
Age when enrolled, y			
18–24	250	164	65.6
25–34	184	120	65.2
35–44	287	194	67.6
P value*			.900
Gender			
Female	450	293	65.1
Male	271	185	68.3
P value			.490
Race/ethnicity			
White	325	206	63.4
Black or African American	318	223	70.1
Asian	19	13	68.4
Hispanic	30	20	66.7
Other or unstated	29	16	55.2
P value			.978
Study site			
Baltimore, MD	303	184	60.7
Buffalo, NY	153	106	69.3
Chapel Hill, NC	113	87	77.0
Gainesville, FL	152	101	66.5
P value			.009
Risk predictors recorded at baseline			
No. of tender body sites			
None	346	238	68.8
1–3	216	140	64.8
≥4	159	100	62.9
P value			.416
Stroop-Pain Δ MAP, mmHg			
<−2	273	174	63.7
−2 to +2	205	136	66.3
>+2	121	90	74.4
P value			.291
Pressure pain threshold: trapezius, kPa			
<275	244	155	63.5
275–<440	240	170	70.8
≥440	227	145	63.9
P value			.671
Perceived Stress Scale			
<12	168	105	62.5
12–<18	243	159	65.4
≥18	307	212	69.1
P value			.071
Risk predictors recorded in quarterly health update			
Number of nonspecific orofacial symptoms			
None	188	134	71.3
1	210	133	63.3
2	162	116	71.6
≥3	161	95	59.0
P value			.109
Time since enrollment, mo			
<12	180	143	79.4
11–<20	194	137	70.6
20–<30	163	113	69.3
≥30	184	85	46.2
P value			<.001

**P* values are from score-statistic for Type III generalized estimating equation analysis of null hypothesis that percentage examined is equivalent among subgroups. The generalized estimating equation model adjusted for clustering of symptom episodes within people.

symptomatic episodes that were not examined as intended, the incidence rate increased to 3.9% per annum (Table 5). Most of the associations seen with the complete case analysis were similar in this imputed analysis, with hazard ratios differing by no more than .1 in absolute value. Among the baseline risk factors, the largest difference was observed for the highest category of body tenderness, where the imputed hazard ratio of 1.48 (95% CLs = 1.10, 1.99) was lower than the complete-case hazard ratio of 1.77 (95% CLs = 1.32, 2.37). For the autonomic measure, the imputed hazard ratio was farther from the null compared to the complete-case hazard ratio, whereas for 1 category of pressure pain thresholds, the imputed hazard ratio was closer to the null compared to the complete-case hazard ratio.

Single Imputation of Incidence to Account for False-Negative Rate of Asymptomatic Quarterly Health Updates

Although the false-negative rate of quarterly health updates was low, it was greater than zero. Although there were too few false negatives to model asymptomatic events, it seemed unlikely that the probability of false negatives would be equivalent for successive, asymptomatic events. This supposition was confirmed in univariate analysis of the data from 338 asymptomatic examinations deemed dependable (ie, excluding examiner 4): 107 of those examinations were from people who had only 1 asymptomatic episode, whereas 180 examinations were from 90 people who reported 2 asymptomatic episodes. (The remaining 51 asymptomatic examinations were of participants who also had symptomatic examinations.) The observed probability of clinically classified TMD was 11.2% for participants with a single asymptomatic episode but only 4.4% for participants with 2 asymptomatic episodes. The latter is much lower than the expected probability of 21.1% that would occur according to the binomial theorem if the probability of a false positive in the second asymptomatic episode were equal to and independent of the probability for a single asymptomatic episode.

The approach to single imputation therefore applied the observed probability of a single false negative to the 190 participants who reported only a single, asymptomatic episode during follow-up but who did not have a follow-up examination by design. For the remaining

Table 3. Examination Outcomes Following Quarterly Health Updates: OPPERA Prospective Cohort Study, 2006–2011

	Symptomatic Examinations*		Asymptomatic Examinations	
	Number of Questionnaires	% Classified as TMD (PPV)	Number of Questionnaires	% Classified as Non-TMD (NPV)
Demographics and study site				
All people	478	49.2	338	92.6
Age when enrolled, y				
18–24	164	57.9	151	90.1
25–34	120	60.0	89	92.1
35–44	194	35.1	98	96.9
P value†		<.001		.066
Gender				
Female	293	51.2	217	91.7
Male	185	45.9	121	94.2
P value		.326		.385
Race/ethnicity				
White	206	58.7	183	90.2
Black or African American	223	37.2	86	96.5
Asian	13	69.2	37	100.0
Hispanic	20	80.0	25	88.0
Other or unstated	16	37.5	7	85.7
P value		<.001		nc
Study site				
Baltimore, MD	184	28.8	95	95.8
Buffalo, NY	106	61.3	102	94.1
Chapel Hill, NC	87	32.2	46	95.7
Gainesville, FL	101	88.1	95	86.3
P value		<.001		.137
Risk predictors recorded at baseline				
No. of tender body sites				
None	238	35.7	154	96.8
1–3	140	50.7	92	92.4
≥4	100	79.0	92	85.9
P value		<.001		.016
Stroop-Pain Δ MAP, mmHg				
<−2	174	49.4	93	91.4
−2 to +2	136	52.9	103	90.3
>+2	90	51.1	79	93.7
P value		.871		.698
Pressure pain threshold: trapezius, kPa				
<275	155	47.7	122	92.6
275–<440	170	53.5	99	91.9
≥440	145	45.5	114	93.9
P value		.445		.856
Perceived Stress Scale				
<12	105	57.1	118	94.1
12–<18	159	49.1	135	93.3
≥18	212	45.3	83	89.2
P value		.172		.469
Risk predictors recorded at QHU				
Number of nonspecific orofacial symptoms				
None	134	39.6	273	95.6
1	133	38.3	33	87.9
2	116	56.0	18	94.4
≥3	95	69.5	14	42.9
P value		<.001		.022
Time since enrollment, mo				
<12	143	60.8	69	95.7
11–<20	137	53.3	93	91.4
20–<30	113	39.8	88	89.8
≥30	85	35.3	88	94.3
P value		<.001		.441

Table 3. Continued

	Symptomatic Examinations*		Asymptomatic Examinations	
	Number of Questionnaires	% Classified as TMD (PPV)	Number of Questionnaires	% Classified as Non-TMD (NPV)
Examiner at follow-up				
Examiner 4	75	100.0	34	76.5
Other examiners at same site as 4	24	50.0	60	91.7
Examiners at other sites	379	39.1	244	95.1
P value		nc		.049

NOTE. nc = *P* value not calculated because of 0% or 100% cell.
*Symptomatic examinations were those conducted following a quarterly health update in which orofacial pain symptoms were reported. Asymptomatic examinations were those conducted following a quarterly health update in which **no** orofacial symptoms were reported. PPV = positive predictive value of the quarterly health update screening questions regarding orofacial pain. NPV = negative predictive value of the same screening questions.
†*P* values are from score-statistic for Type III generalized estimating equation analysis of null hypothesis that examination outcome is equivalent among subgroups. The generalized estimating equation model adjusted for clustering of examinations within people.

2,023 participants reporting 2 or more asymptomatic episodes and who did not did not have a follow-up examination, a 4.4% probability of false negatives was assumed. (The 534 participants who reported both asymptomatic and symptomatic episodes during follow-up were excluded from these calculations on the grounds that their symptomatic episodes initiated a follow-up examination, thereby allowing for an observed event of TMD.) This yielded an expected number of 110 participants with false-negative asymptomatic episodes. When these cases were added to the 260 observed cases, the imputed incidence rate increased to 5.0% per annum.

Table 4. Interexaminer Reliability* of 7 Examiners: OPPERA Prospective Cohort Study, 2006–2011

Date	Examiner	Number of Paired Examinations	Kappa (95% CLs)
February 2006	7	24	.83 (.60, 1.00)
	4	24	.92 (.74, 1.00)
	5	24	.92 (.74, 1.00)
April 2006	5	16	1.00 (n/a)†
	6	16	1.00 (n/a)
October 2006	4	15	1.00 (n/a)
May 2007	7	15	1.00 (n/a)
June 2007	7	16	1.00 (n/a)
	4	16	1.00 (n/a)
	5	16	1.00 (n/a)
	6	16	1.00 (n/a)
February 2009	7	20	1.00 (n/a)
	4	20	1.00 (n/a)
	2	19	.82 (.24, 1.00)
	6	20	1.00 (n/a)
June 2010	7	20	.89 (.74, 1.00)
	3	20	.89 (.77, 1.00)
	6	20	.89 (.58, 1.00)
	1	15	1.00 (n/a)
October 2011	7	20	1.00 (n/a)
	1	20	1.00 (n/a)
	3	20	1.00 (n/a)
	6	20	1.00 (n/a)

*Reliability of TMD case classification was assessed in paired examinations of volunteers who were not study participants: one of the paired examinations was conducted by the OPPERA study-site examiner and the other paired examination was conducted by the OPPERA reference examiner.
†95% CLs are not applicable for kappa values of 1.00.

Discussion

Despite adopting rigorous follow-up procedures in this prospective cohort study, participants were lost to follow-up and methods used to enumerate cases of first-onset TMD were imperfect. Furthermore, quality assessment of examination data suggested that TMD case classifications recorded by 1 examiner were dubious, notwithstanding extensive training and calibration and excellent interexaminer reliability. Conventional methods of complete case analysis, which use only the observed data, ignore potential biases created by these problems. Hence, 2 analytic strategies were used to assess the degree of bias. Hot-deck multiple imputation, which accounted for complete loss to follow-up, yielded identical overall incidence and generally similar or slightly attenuated hazard ratios compared to complete-case analysis, suggesting that loss to follow-up was not a serious source of bias. A 2-stage, multiple-imputation procedure addressed problems arising from nonexamination—or from dubious examination—of symptomatic episodes. The overall annual incidence rate of 3.9% was slightly greater than the rate of 3.5% from the complete-case analysis, and although differences in hazard ratios were small, both attenuation and amplification was observed. The problem of false negatives in follow-up questionnaires could be evaluated only using single imputation, with the results suggesting that the true incidence rate in the cohort might be as high as 5.0% per annum. Taken as a whole, these findings suggest that a small, though not ignorable, amount of bias was created by missing or dubious data from follow-up. Although the bias did not appreciably alter "bottom line" statistical conclusions for the few risk factors investigated here, it is prudent to consider multiply-imputed findings when evaluating the large number of risk factors under investigation in this study.

Study participants' duration of follow-up in this study varied not only because of loss to follow-up, but also because of the study design: a single census date was

Table 5. Sensitivity Analysis of TMD Incidence Rates and Hazard Ratios: OPPERA Prospective Cohort Study, 2006–2011

	Complete Case Analysis			Imputation for Loss to Follow-Up			Imputation for People Who Were Not Examined as Intended		
Group	No. of People	Incidence Rate	HR (95% CLs)	No. of People	Incidence Rate	HR (95% CLs)	No. of People	Incidence Rate	HR (95% CLs)
All people	2,737	3.5	3.2, 3.9	3,258	3.5	3.2, 3.9	2,737	3.9	3.5, 4.3
Age when enrolled, y									
18–24	1,421	2.9		1,706	3.0		1,421	3.1	
25–34	736	3.8	1.34 (1.01, 1.79)	860	3.8	1.28 (.97, 1.70)	736	4.1	1.32 (.98, 1.78)
35–44	580	4.7	1.66 (1.23, 2.24)	692	4.6	1.56 (1.17, 2.09)	580	5.5	1.76 (1.30, 2.39)
Gender									
Female	1,107	3.1		1,396	3.3		1,107	3.3	
Male	1,630	3.8	1.21 (.94, .15)	1,862	3.7	1.13 (.88, .33)	1,630	4.2	1.27 (.97, .08)
Race/ethnicity									
White	1,448	3.3		1,637	3.4		1,448	3.7	
Black/African American	766	4.7	1.39 (1.06, 1.81)	1,012	4.5	1.32 (1.01, 1.72)	766	5.4	1.48 (1.13, 1.94)
Asian	256	1.4	.40 (.20, .78)	299	1.6	.46 (.24, .88)	256	1.6	.42 (.21, .85)
Hispanic	178	3.9	1.16 (.72, 1.88)	211	3.8	1.12 (.70, 1.80)	178	3.4	.90 (.51, 1.61)
Other or unstated	89	2.8	.87 (.41, 1.87)	99	2.9	.89 (.42, 1.88)	89	3.2	.86 (.40, 1.82)
Study site									
Buffalo, NY	693	3.9		797	3.9		693	4.9	
Baltimore, MD	574	4.2	1.07 (.75, 1.51)	768	4.2	1.07 (.76, 1.50)	574	5.3	1.07 (.76, 1.48)
Gainesville, FL	765	4.8	1.25 (.93, 1.70)	878	4.5	1.21 (.89, 1.63)	765	4.2	.83 (.61, 1.15)
Chapel Hill, NC	705	1.4	.38 (.25, .58)	815	1.7	.46 (.30, .70)	705	1.8	.36 (.24, .54)
No. of tender body sites									
None	1,301	2.6		1,584	2.8		1,301	3.1	
1–3	682	4.1	1.62 (1.20, 2.20)	808	4.1	1.50 (1.12, 2.01)	682	4.5	1.43 (1.05, 1.95)
≥4	754	4.5	1.77 (1.32, 2.37)	866	4.4	1.62 (1.22, 2.16)	754	4.6	1.48 (1.10, 1.99)
Stroop-Pain Δ MAP, mmHg									
<−2	814	4.4		982	4.2		814	4.9	
−2 to +2	774	3.9	.90 (.67, 1.21)	911	3.9	.91 (.68, 1.22)	774	4.2	.85 (.63, 1.15)
>+2	625	2.9	.67 (.48, .94)	716	2.9	.70 (.50, .97)	625	2.9	.59 (.41, .86)
Pressure pain threshold: trapezius, kPa									
<275	922	2.9		1,057	3.0		922	3.3	
275–<440	897	4.0	1.39 (1.03, 1.88)	1,089	4.0	1.33 (.99, 1.79)	897	4.1	1.23 (.90, 1.69)
≥440	895	3.4	1.18 (.86, 1.62)	1,082	3.5	1.18 (.86, 1.61)	895	4.1	1.22 (.89, 1.67)
Perceived Stress Scale									
<12	914	2.5		1,035	2.7		914	2.9	
12–<18	962	3.3	1.27 (.93, 1.75)	1,132	3.3	1.25 (.91, 1.71)	962	3.6	1.25 (.90, 1.73)
≥18	847	5.1	1.96 (1.44, 2.66)	1,064	4.8	1.77 (1.31, 2.40)	847	5.5	1.97 (1.43, 2.70)

used for participants enrolled over a period of more than 2 years. A conventional method to account for differential periods of follow-up is to express incidence as the rate of occurrence, rather than the proportion of people who develop the condition. The incidence rate is also labeled the "force of morbidity"[20] because it considers both the development of a condition and the rapidity of onset. Furthermore, incidence rate ratios provide an intuitive measure of the magnitude of association between hypothesized risk factors and TMD incidence. Hazard ratios provide a good approximation of the incidence rate ratio and can be modeled, with few assumptions, using Cox regression models. In this analysis, the principal statistical assumption of proportional hazards was satisfied for the large number of putative risk factors assessed.

A more serious concern is the possibility that loss to follow-up might occur differentially according to putative risk factors, thereby biasing estimates. The 42% rate of complete follow-up in this study was within the range reported for other, population-based studies that have used multiple follow-up assessments to enumerate incidence of pain. In the UK study of back pain, 30% of enrollees completed all 6 of the intended follow-up questionnaires,[8] whereas in the U.S. study of 4 types of pain in adolescents, 64% completed all 11 of the intended follow-up questionnaires.[9] In principle, associations from complete-case analysis of such data are more biased than estimates from imputed data sets that appropriately account for missing data.[22] Multiple imputation is a favored method for many types of data, although imputation of the outcome measure in

time-to-event data creates particular challenges when the data are to be analyzed with Cox regression models,[4] as in this setting.

The 2-stage method of multiple imputation was developed to deal with those challenges. When evaluated in simulation studies, the method was markedly superior to complete case analysis with respect to bias, coverage, and confidence interval width under various assumptions regarding the pattern of missing data and characteristics of variables used for imputation.[2] Using the data from this cohort, the method produced slightly greater estimates of the incidence rate than the complete-case analysis. Differences in hazard ratios were likewise small, although compared to the complete-case analysis, the estimates moved in varying directions, either toward or away from the null, according to the risk factor being analyzed. This is consistent with the findings of an earlier study by Cook and Kosorok[4] who found that there is little loss in efficiency when up to 50% of possible events in a prospective cohort study are unadjudicated. Nevertheless, both the simulation findings and the current sensitivity analysis indicated some bias in estimates from complete-case analysis, which, though usually small in magnitude, should not be ignored when evaluating putative risk factors. For that reason, accompanying papers in this issue report both unimputed (complete-case) and imputed findings that account for missing or dubious data from follow-up examinations.

The problem of dubious case classifications by 1 examiner was unexpected given the intensity of examiner training conducted prior to study commencement and annually thereafter. Furthermore, the examiner and the study's reference examiner had excellent levels of agreement in annual studies of interexaminer reliability. Problems were first suspected during annual data quality assessment, and the findings were reported to the OPPERA Scientific Advisory Committee in December 2008. The Committee advised the investigators to investigate the problem further using more detailed information as it accrued. By the time the problem was confirmed and reported to the Committee at its December 2009 meeting, the examiner had stopped working in the study. The Committee then advised the investigators to develop strategies for analyses that acknowledged and addressed the problem. Accordingly, this paper has focused on the problem and used the second method of multiple imputation[2] to address it.

As noted above, frequent follow-up assessments are needed to properly enumerate incidence of pain over a period of years. Because reexamination at that frequency for all subjects is not feasible in population studies, high-quality screening mechanisms are required. The questions in the quarterly health updates that screened for orofacial pain had excellent test-retest reliability. Based on the combination of sensitivity (90%) and specificity (56%), they also had good validity that exceeded a previous study of adults.[23] In formulating the screening questions, greater emphasis was given to specificity than to sensitivity on the grounds that false positives (which reduce sensitivity) would be identified during examination and therefore not counted as incident cases. Nonetheless, future studies should consider using a new screening questionnaire for TMD, developed after commencement of OPPERA, which has better sensitivity and specificity.[13]

Despite the small proportion of false negatives in this study, it was not ignorable because the majority of participants reported no symptoms throughout the study. Indeed, the estimated incidence rate increased to 5% per annum with single imputation to account for the false negatives. Ideally, multiple imputation of these false-negative events would allow sensitivity analysis to determine the effects on hazard ratios. However, there were too few false negatives to permit such analysis.

Although loss to follow-up in prospective cohort studies is potentially a serious threat, the study design has inherent strengths for etiologic research, including its focus on prognosis and its determination of a temporal association between putative causes and effects. Furthermore, prospective cohort studies are less prone to selection bias, which is a particular problem in case-control studies of pain.[26] For example, if a case-control study recruited potential cases from among patients seeking health care, whereas controls were recruited from the community, cases likely would have greater odds of health insurance coverage. The resulting odds ratio would spuriously suggest that recent health insurance increased the risk of TMD. Furthermore, other spurious associations probably would be seen for characteristics associated with health insurance. In contrast, prospective cohort studies enumerate TMD during follow-up examinations using the same procedures for all enrollees, regardless of baseline characteristics such as health insurance. This independence between putative risk factors and procedures used in enumerating cases vastly diminishes the potential for selection bias, which is a major concern in case-control studies.

Results from this study can be generalized only to the target population of 18- to 44-year-olds with no significant history of TMD symptoms and no serious health conditions. The age range is one in which U.S. population surveys[19,30,32] show an age-associated increase in prevalence, peaking in the fifth decade. It effectively excludes postmenopausal women whose risk of TMD likely is influenced by changes in reproductive hormones.[21] In a separate paper in this issue,[33] we discuss the implications of this age restriction when comparing OPPERA results with other cohort studies. Another caveat is that participants were not a probability sample drawn at random from the population, but rather volunteers recruited by advertisements at 4 U.S. study sites. Elsewhere,[32] we have reported that the distribution of sociodemographic characteristics in this OPPERA cohort is similar to some, although not all, benchmarks from the U.S. population census. We therefore concluded that findings are broadly applicable to major demographic groups in the United States.

Another important caveat about generalizability is that this study focused on first-onset TMD in an initially symptom-free cohort. Recurrent bouts of TMD were not

enumerated, and hence the total burden of TMD in the population was underestimated. Conversely, though, some participants might have been enrolled incorrectly, having forgotten about episodes of TMD experienced many years before they were screened for eligibility. Indeed, in a separate study of TMD symptoms in this cohort,[34] we found that one third reported orofacial pain for ≥5 days per month in at least 1 month during follow-up and we speculated that some individuals likely had forgotten about self-limiting episodes of pain that occurred many years prior to enrollment in a study.

In summary, this paper has documented 3 critical methodological components of the OPPERA prospective cohort study: 1) mindful of the intermittent nature of pain, the study was designed with quarterly follow-up assessments to optimize enumeration of first-onset TMD; 2) particular attention was paid to data quality, including annual examiner training and procedures to monitor reliability and validity of the study's main outcomes; and 3) rigorous statistical methods were adapted to evaluate and reduce potential biases that can occur because of the inevitable problem of missing follow-up data. The findings support the study's validity for the purpose of investigating the etiology of first-onset TMD and provide the foundation for other papers investigating risk factors hypothesized in the OPPERA project.

Acknowledgments

The authors thank the OPPERA research staff for their invaluable contributions to this work. In addition, we express our gratitude to the participants who have devoted time and effort in support of this research. This work was done at University of North Carolina at Chapel Hill, NC; University at Buffalo, NY; University of Maryland–Baltimore, MD; University of Florida, FL; and Battelle Memorial Institute, NC.

Supplementary Data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jpain.2013.06.006.

References

1. Andridge RR, Little RJ: A review of hot deck imputation for survey non-response. Int Stat Rev 78:40-64, 2010

2. Brownstein N, Cai J, Slade G, Bair E: Parameter estimation in Cox proportional hazard models with missing censoring indicators. Available at: http://arxiv.org/abs/1304.3839. Accessed April 18, 2013

3. Carey TS, Garrett JM, Jackman A, Hadler N: Recurrence and care seeking after acute back pain: Results of a long-term follow-up study. North Carolina Back Pain Project. Med Care 37:157-164, 1999

4. Cook TD, Kosorok MR: Analysis of time-to-event data with incomplete event adjudication. J Am Stat Assoc 99: 1140-1152, 2004

5. Croft P, Dunn K, Blyth FM, Windt Dvd: Definition and measurement of chronic pain for populations, in Croft P, Blyth FM, Windt Dvd (eds): Chronic Pain Epidemiology: From Aetiology to Public Health. Oxford, Oxford University Press, 2010, pp 37-43

6. Diatchenko L, Slade GD, Nackley AG, Bhalang K, Sigurdsson A, Belfer I, Goldman D, Xu K, Shabalina SA, Shagin D, Max MB, Makarov SS, Maixner W: Genetic basis for individual variations in pain perception and the development of a chronic pain condition. Hum Mol Genet 14: 135-143, 2005

7. Drangsholt M, LeResche L: Temporomandibular disorder pain, in Crombie IK, Croft PR, Linton SJ, LeResche L, Von Korff M (eds): Epidemiology of Pain. Seattle, WA, IASP Press, 1999, pp 43-52

8. Dunn KM, Jordan K, Croft PR: Characterizing the course of low back pain: A latent class analysis. Am J Epidemiol 163:754-761, 2006

9. Dunn KM, Jordan KP, Mancl L, Drangsholt MT, Le Resche L: Trajectories of pain in adolescents: A prospective cohort study. Pain 152:66-73, 2011

10. Dworkin S, LeResche L: Research diagnostic criteria for temporomandibular disorders: Review, criteria, examinations and specifications, critique. J Craniomandib Disord 6: 301-355, 1992

11. 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 14:T75-T90, 2013

12. 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 12:T46-T60, 2011

13. Gonzalez YM, Schiffman E, Gordon SM, Seago B, Truelove EL, Slade G, Ohrbach R: Development of a brief and effective temporomandibular disorder pain screening questionnaire: Reliability and validity. J Am Dent Assoc 142:1183-1191, 2011

14. Grambsch PM, Therneau TM: Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81: 515-526, 1994

15. 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 14:T63-T74.e6, 2013

16. 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 12:T61-T74, 2011

17. Haythornwaite JA, Fauerbach JA: Assessment of acute pain, pain relief, and patient satisfaction, in Turk DC, Melzack R (eds): Handbook of Pain Assessment, 2nd ed. New York, NY, Guilford Press, 2001, pp 417-430

18. Hill AB: The environment and disease: Association or causation? Proc R Soc Med 58:295-300, 1965

19. Isong U, Gansky S, Plesh O: Temporomandibular joint and muscle disorder-type pain in U.S. adults: The National Health Interview Survey. J Orofac Pain 22: 317-322, 2008

20. Last JM, Spasoff RA, Harris SS, Thuriaux MC, International Epidemiological Association: A Dictionary of Epidemiology, 4th ed. Oxford, Oxford University Press, 2001

21. LeResche L: Epidemiology of temporomandibular disorders: Implications for the investigation of etiologic factors. Crit Rev Oral Biol Med 8:291-305, 1997

22. Little RJA, Rubin DB: Statistical Analysis With Missing Data, 2nd ed. Hoboken, NJ, Wiley, 2002

23. Locker D, Slade G: Association of symptoms and signs of TM disorders in an adult population. Community Dent Oral Epidemiol 17:150-153, 1989

24. 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 12, 2011. T4–11. e1–T4–11.e2

25. 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 12:T75-T91, 2011

26. Marbach JJ, Schwartz S, Link BG: The control group conundrum in chronic pain case/control studies. Clin J Pain 8:39-43, 1992

27. Neogi T, Nevitt MC, Yang M, Curtis JR, Torner J, Felson DT: Consistency of knee pain: Correlates and association with function. Osteoarthritis Cartilage 18:1250-1255, 2010

28. Ohrbach R, Bair E, Fillingim RB, Gonzalez Y, Gordon SM, Lim P-F, 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 14:T33-T50, 2013

29. 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 12: T27-T45, 2011

30. Plesh O, Adams SH, Gansky SA: Racial/ethnic and gender prevalences in reported common pains in a national sample. J Orofac Pain 25:25-31, 2011

31. Rothman KJ, Greenland S, Lash TL: Modern Epidemiology, 3rd ed. Philadelphia, PA, Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008

32. 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 12:T12-T26, 2011

33. 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 14:T20-T32.e3, 2013

34. Slade GD, Sanders AE, Bair E, Brownstein N, Dampier D, Knott C, Fillingim R, Maixner WO, Smith S, Greenspan J, Dubner R, Ohrbach R: Preclinical episodes of orofacial pain symptoms and their association with health care behaviors in the OPPERA prospective cohort study. Pain 154:750-760, 2013

35. Smith SB, Maixner DW, Greenspan JD, Dubner R, Fillingim RB, Ohrbach R, Knott C, Slade GD, Bair E, Gibson DG, Zaykin DV, Weir BS, Maixner W, Diatchenko L: Potential genetic risk factors for chronic TMD: Genetic associations from the OPPERA case control study. J Pain 12: T92-T101, 2011

36. Soni A, Kiran A, Hart DJ, Leyland KM, Goulston L, Cooper C, Javaid MK, Spector TD, Arden NK: Prevalence of reported knee pain over twelve years in a community-based cohort. Arthritis Rheum 64:1145-1152, 2012

37. van Oostrom SH, Monique Verschuren WM, de Vet HC, Picavet HS: Ten year course of low back pain in an adult population-based cohort—The Doetinchem cohort study. Eur J Pain 15:993-998, 2011

38. Von Korff M: Studying the natural history of back pain. Spine (Phila Pa 1976) 19:2041S-2046S, 1994