

ARIC Manuscript Proposal #2276

PC Reviewed: 12/10/13
SC Reviewed: _____

Status: A
Status: _____

Priority: 2
Priority: _____

1. a. Full Title: Estimating full-mouth prevalence of dental diseases from partial mouth periodontal examination using a statistical model for correlated binary data

b. Abbreviated Title (Length 26 characters): Dental Prevalence Estimates

2. Writing Group:

Writing group members: Sarah Marks, John Preisser, Anne Sanders, Jim Beck, others (?)

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. SJM [**please confirm with your initials electronically or in writing**]

First author: Sarah Marks

Address: Department of Biostatistics, CB#7420
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420

Phone: 404-313-1903

Fax: 919-966-3804

E-mail: sjmarks@live.unc.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

Name: John Preisser

Address: Department of Biostatistics
Department of Biostatistics, CB#7420
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420

Phone: 919-966-7265

Fax: 919-966-3804

E-mail: john_preisser@unc.edu

3. Timeline: Analyses to commence immediately following committee approval; draft manuscript complete by mid-April for submission as a UNC Department of Biostatistics master's paper. Goal is submission of MS to a journal before June 1, 2014.

4. Rationale:

The gold standard assessment for estimating prevalence of dental conditions such as periodontitis and dental caries is a full mouth examination. A full mouth exam involves inspection of up to 168 dental sites: six sites on 28 teeth (3rd molars are usually excluded). These examinations are time intensive, requiring 25 to 45 minutes per exam, and thus are costly and impractical for purposes of research and surveillance. Consequently partial mouth recording protocols (PRPs) are often used in large scale epidemiological studies and dental surveys instead of full mouth exams; PRPs include both random site selection methods (RSSMs) based on simple random samples and fixed site selection methods (FSSMs), where either specific teeth or sites or both are selected.

PRPs significantly underestimate prevalence of dental conditions in a population. As disease classification is commonly based on the presence of the condition at one or more dental sites, sampling methods inherently lead to underestimates of the proportion of individuals affected by dental conditions when only a subset of sites is examined. Previous research has quantified the extent to which PRP techniques underestimate prevalence (Beck et al. 2006). Using probing depths of greater than 4, 5 or 6 mm and clinical attachment levels greater than 3, 4, 5 or 6 mm on at least one tooth as indicators of periodontitis, Beck et al. found that whole-mouth prevalence was underestimated anywhere from 1 to 78% depending on the type of sampling method used and the probing depth/clinical attachment level selected as a cut-point. All FSSMs showed greater underestimation than RSSMs for comparable number of sites. Thus, there is a need to develop innovative statistical methods to provide unbiased estimation of prevalence to validate partial mouth sampling methods and capitalize on their strengths. The proposed manuscript will introduce new estimators of prevalence for RSSMs, and apply and evaluate these using the ARIC Visit 4 dental exam data (ARIC ancillary study #1996.01, principal investigator: Beck JD).

5. Main Hypothesis/Study Questions:

To overcome the current limitations of simple estimators for PRPs, we propose a statistical model for binary correlated data to estimate the prevalence from partial recording. Specifically, new estimators of prevalence for RSSMs will be based on the conditional linear family of models for correlated binary distributions (Qaqish et. al. 2003).

Specific aim 1: The proposed method to estimate the prevalence of periodontitis from RSSMs will be applied using several clinical attachment levels and probing depth cut-points. These estimates will be compared with prevalence estimates based on full mouth exams for the whole ARIC cohort having a dental exam and for several subpopulations.

We hypothesize that the new estimates for the dental ARIC cohort based on RSSMs will be very similar to the gold standard estimates based on full mouth recording. Bootstrap standard errors for the prevalence estimates will be developed and illustrated.

Specific aim 2: The accuracy of the new prevalence estimates based on RSSMs will be assessed with a simulation study based on repeated sampling from the ARIC full-mouth exam multivariate Bernoulli data. We hypothesize that the new estimates will have negligible bias with respect to the gold standard estimates based on full mouth recording, and that the precision of the new estimator will increase as the number of sites selected with RSSM protocols increases.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

We will first develop a new statistical model to use PRPs to obtain accurate estimates of prevalence of dental conditions such as periodontitis. We will then evaluate this model using dental ARIC data. Specifically, we will treat periodontitis as a subject level binary variable; as an example, we might treat any site with a probing pocket depth of ≥ 4 mm as indicative of periodontitis. Our goal is to estimate the proportion of individuals in a population that has at least one site with a probing depth of ≥ 4 mm. To estimate this, we start by considering an individual with all teeth remaining and with up to probing 168 sites. Since each site is from the same individual, we assume that there is a correlation between sites: individuals who have periodontitis at one site are more likely to have periodontitis at another site relative to periodontitis from a site of another individual. We will denote each site as Y_j where $j = 1, \dots, n$ and $n = 168$ for our model. We can denote the j th site as having periodontitis by $Y_j = 1$ or not having periodontitis by $Y_j = 0$; thus, we can say an individual has periodontitis when at least one of their 168 sites has periodontitis or $\sum_{j=1}^{168} Y_j \geq 1$. Since prevalence is the total number of people in a given population with periodontitis divided by the population at risk, this is equivalent to the probability that a randomly selected individual in the population has periodontitis. Using our previous definition of periodontitis on an individual level, we can thus write the population prevalence, π , as $\pi = P(\sum_{j=1}^{168} Y_j \geq 1)$. We can also think of this population prevalence as one minus the probability that all sites do not have periodontitis or $\pi = 1 - P(\sum_{j=1}^{168} Y_j = 0) = 1 - P(Y_1 = 0, \dots, Y_n = 0)$.

Comment [AES1]: edentulous people are excluded from the denominator

To estimate this population prevalence, we will use a statistical model based on the conditional linear family (CLF) of correlated Bernoulli distributions (Qaqish, 2003). Use of CLF requires specification of the marginal means (probabilities) $\mu_j = E(Y_j) = P(Y_j)$ for $j = 1, \dots, n$ and $\rho_{j,k}$, the pairwise correlation coefficient between any two sites $\rho_{j,k} = \text{Corr}(Y_j, Y_k)$ for $j = 1, \dots, n - 1$ and $k = j + 1, \dots, n$. For a simple working model, we assume that all sites have the same probability of having periodontitis ($\mu_j = \mu, j = 1, \dots, n$) and the correlation between having periodontitis at any two sites is the same $\rho_{j,k} = \rho$ for $j = 1, \dots, n - 1$ and $k = j + 1, \dots, n$. Although these assumptions

are most likely not true, it is hypothesized that it should produce an estimate of prevalence with minimum bias because with a RSSM the sites are randomly selected. The working model leads to the following CLF expression for prevalence $\pi = 1 - (1 - \mu)\prod_{j=2}^n \left(1 - \frac{(1-\rho)\mu}{1+(j-2)\rho}\right)$. In order to estimate π , we plug-in estimates of μ and ρ , obtained with simple estimators, the latter a GEE-type method-of-moments exchangeable correlation estimator. Other working models would lead to different CLF distributions (Preisser and Qaqish 2014), which would then give different estimators of prevalence.

We will evaluate the estimator described above using data from ARIC. During visit 4, 6793 ARIC participants had a full periodontal assessment. This enables us to simulate RSSM PRPs by drawing samples from each participant, and then to estimate prevalence using the newly developed statistical model. We can compare these prevalence estimates from our model with the gold standard of the full mouth exam. We will apply the sample sizes (number of sites) and cut-points for periodontitis as used by Beck et al. (2006). Specifically, we will use the newly developed model with sample sizes of 6, 10, 15, 20, 28, 36, 42 and 84 sites per mouth to estimate the prevalence of periodontitis. We will use the same cut-points for assessing periodontitis, defining the prevalence of periodontitis based on the proportion of people with one or more 4, 5, and 6 mm probing depths and 3, 4, 5, and 6 mm clinical attachment levels. Prevalence will be estimated using the proposed method for the overall ARIC cohort and for specific subpopulations based on gender, race, education, income, smoking, Type 2 diabetes and dental visits (regular vs. episodic); the categories and definitions of these groups are contained in Beck et al. 2006.

To rigorously assess the validity of the model to provide full mouth estimates, we will use repeated sampling of a fixed number of sites per mouth from the ARIC data to estimate means and their standard errors for the prevalence estimates given by our model. It is hypothesized that the new estimates will have negligible bias with respect to the gold standard estimates based on full mouth recording, and that the precision of the new estimator will increase as the number of sites selected with RSSM protocols increases.

References

Beck JS, Caplan DJ, Preisser JS, Moss K. Reducing the bias of probing depth and attachment level estimates using random partial-mouth recording. *Community Dentistry and Oral Epidemiology* (2006) 34:1-10

Preisser JS, Qaqish BF. A comparison of methods for simulating correlated binary variables with specified marginal means and correlations. *Journal of Statistical Computation and Simulation* (2014). (published online 15 Jul 2013). DOI: 10.1080/00949655.2013.818148

Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* (2003) 90:455-463.

7.a. Will the data be used for non-CVD analysis in this manuscript? Yes
 No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = "CVD Research" for non-DNA analysis, and for DNA analysis RES_DNA = "CVD Research" would be used?

Yes No

(This file ICTDER has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript?
 Yes No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = "No use/storage DNA"?

Yes No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: <http://www.csc.unc.edu/ARIC/search.php>

Yes No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

MS #2191. Two of the co-authors of the proposed MS (Preisser, Beck) are co-authors on MS #2191 and will coordinate the two MS.

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? Yes No

11.b. If yes, is the proposal

A. primarily the result of an ancillary study (list number* _1996.01_)

B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* _____)

*ancillary studies are listed by number at <http://www.csc.unc.edu/aric/forms/>

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PUBMED Central** whenever the journal does not and be in compliance with this policy. Four files about the public access policy from <http://publicaccess.nih.gov/> are posted in <http://www.csc.unc.edu/aric/index.php>, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to Pubmed central.