In the current era of diet-gene analyses, large sample sizes are required to uncover the etiology of complex diseases. As such, consortia form and often combine available data. Food frequency questionnaires, which commonly use 2 different types of responses about the frequency of intake (predefined responses and open-ended responses), may be pooled to achieve the desired sample size. The common practice is to categorize open-ended responses into the predefined response categories. A problem arises when the predefined categories are noncontiguous: possible open-ended responses may fall in gaps between the predefined categories. Using simulated data modeled from frequency of intake among 1,664 controls in a lung cancer case-control study at The University of Texas M. D. Anderson Cancer Center (Houston, Texas, 2000–2005), the authors describe the effect of different categories of open-ended responses that fall in between noncontiguous, predefined response sets on estimates of the mean difference in intake and the odds ratios. A significant inflation of false positives appears when comparing mean differences of intake, while the bias in estimating odds ratios may be acceptably small. Therefore, if pooling data cannot be restricted to the same type of response, inferences should focus on odds ratio estimation to minimize bias.

In the life span of long-running case-control accrual or cohort studies, new technologies for data collection, such as the OpScan dual NSIGHT scanner (Scantron Corporation, Eagan, Minnesota), may lead to changes in study design or data collection methods. Such is true in our own experience. We attempted to redesign a nutrition questionnaire to be more readily processed by converting a questionnaire soliciting open-ended frequency responses to a more easily scanned format using predefined categories. We were then faced with how to pool the older and newer data. This same scenario will also arise when collaborators want to pool their data for a larger sample size where some in the collaboration use an open-ended questionnaire, such as the Block food frequency questionnaire (FFQ) (1), and others use a questionnaire with predefined responses, such as the Willett FFQ (2). When researchers pool dietary data based on both predefined and open-ended responses regarding frequency of intake to study an outcome of interest, they commonly categorize the open-ended responses into the predefined response categories (2). However, this approach has methodological implications for diet-disease risk assessment.

Research has focused on the advantages and disadvantages of using predefined categories versus open-ended responses, ranging from loss of information by categorizing (2–8) to open-ended responses being conducive to error without interviewer assistance (9). (For further discussion on categorical vs. open-ended responses, refer to Subar et al. (9) and Willett (10).) The current paper investigates the error introduced by pooling previously collected open-ended and predefined categorical data. Consider the question, How often did you eat hamburgers?; predefined responses might include the following: less than once per month, 1–3 times per month, once per week, more than once per week, and so on. The corresponding open-ended response for the same

Practice of Epidemiology

Pooling Dietary Data Using Questionnaires With Open-ended and Predefined Responses: Implications for Comparing Mean Intake or Estimating Odds Ratios

Michael D. Swartz*, Michele R. Forman, Somdat Mahabir, and Carol J. Etzel
* Correspondence to Dr. Michael D. Swartz, Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Boulevard, Unit 1340, Houston, TX 77030 (e-mail: mdswartz@mdanderson.org).

Initially submitted July 20, 2009; accepted for publication December 9, 2009.

In the current era of diet-gene analyses, large sample sizes are required to uncover the etiology of complex diseases. As such, consortia form and often combine available data. Food frequency questionnaires, which commonly use 2 different types of responses about the frequency of intake (predefined responses and open-ended responses), may be pooled to achieve the desired sample size. The common practice is to categorize open-ended responses into the predefined response categories. A problem arises when the predefined categories are noncontiguous: possible open-ended responses may fall in gaps between the predefined categories. Using simulated data modeled from frequency of intake among 1,664 controls in a lung cancer case-control study at The University of Texas M. D. Anderson Cancer Center (Houston, Texas, 2000–2005), the authors describe the effect of different categories of open-ended responses that fall in between noncontiguous, predefined response sets on estimates of the mean difference in intake and the odds ratios. A significant inflation of false positives appears when comparing mean differences of intake, while the bias in estimating odds ratios may be acceptably small. Therefore, if pooling data cannot be restricted to the same type of response, inferences should focus on odds ratio estimation to minimize bias.

computer simulation; data interpretation, statistical; diet surveys

Abbreviations: CI, confidence interval; FFQ, food frequency questionnaire; LCPS, lung cancer parent study; OR, odds ratio.
question asks participants to provide their own frequency of intake in units of days, weeks, or months.

When analyzing open-ended responses, researchers usually standardize the open-ended data to a specific time frame, such as times per year, which can then be divided by 52 to calculate the frequency of intake per week. If the categories are contiguous, it is straightforward to categorize the open-ended responses. However, if the predefined categorical responses for frequency of intake are not contiguous, issues arise. In an open-ended questionnaire, one person might respond that he or she eats hamburgers 6 times per month. Consider pooling this response with data from another study using the following predefined categories: once per week and 2–3 times per week. The frequency of intake falls in an undefined gap between the predefined, noncontiguous categories. In this case, the researchers could either categorize the response into once per week, categorize the response into 2–3 times per week, or treat the response as missing. Their decision would ultimately influence the distribution of responses and hence possibly affect inferences about the relation between diet and disease risk.

In our experience, the responses of as many as 11% of the individuals in our sample who provided an open-ended response for frequency of intake fell in a gap between predefined responses. For this study, we applied the categories (and gaps) as defined on the Willett FFQ (2) for the Nurses’ Health Study II. Figure 1 is a graphic representation of the categories and gaps discussed in this paper. To our knowledge, the implications of pooling dietary data from tools with open-ended and predefined, noncontiguous responses have not been reported. This paper investigates whether, and to what extent, pooling dietary data from open-ended and predefined responses affects measures of central tendency and estimation of odds ratios. We report findings using simulated data based on previously collected dietary data from healthy individuals.

**MATERIALS AND METHODS**

This research consisted of 2 phases: a simulation study and a pilot project.
Phase 1: simulation development

Subjects. The study data were previously collected from 1,664 healthy individuals recruited between 2000 and 2005 as controls for a lung cancer case-control study at The University of Texas M. D. Anderson Cancer Center in Houston, Texas (11). Briefly, the controls had no previous diagnosis of cancer (except nonmelanoma skin cancer) and were frequency matched to lung cancer patients on the basis of age (±5 years), smoking status, gender, and ethnicity. The lung cancer parent study (LCPS) recruited the healthy controls at the Kelsey-Seybold Clinic, the largest private physician group in Houston; the overall response rate was about 75%.

Dietary and demographic data. Dietary information, demographic characteristics, and smoking history were collected by personal interviews. Trained interviewers administered a 198-item, modified Block FFQ (1) with open-ended responses regarding frequency of intake. The validity of the Block FFQ has been described across various populations (12, 13). The FFQ was modified for the LCPS by adding foods commonly eaten in the greater Houston metropolitan area. Previously, we showed that the estimated intake of several nutrients (such as iron, magnesium, and calcium) and beverages (including tea and coffee) by healthy controls was comparable with that reported by adults who participated in the 1999–2000 National Health and Nutrition Examination Survey (14–16). For this analysis, we focused on the open-ended responses regarding the frequency of intake of orange juice and hamburgers.

We simulated 2 types of data by modeling the distributions of already collected data on the frequency of intake of orange juice and hamburgers among the 1,664 controls of the LCPS to estimate the population parameters for use in the simulations. First, we generated open-ended responses and predefined responses to compare mean intake measured by each response type; second, we generated case-control data to elucidate the effect on the odds ratio of converting open-ended responses to different predefined response categories.

Mean intake simulations. To evaluate the effect of categorizing open-ended responses on comparing mean intakes between 2 groups, we generated data sets consisting of 2 groups with 3 different sample sizes for each pair of groups: \( n = 40 \), \( n = 200 \), and \( n = 1,000 \) individuals in each of the 2 groups. For each fixed sample size, we generated 200 replicates of the 2 groups.

Simulation of open-ended responses. The empirical distribution for the frequency of intake of orange juice in the LCPS data was positively skewed (data not shown). Therefore, we simulated frequency of intake using an exponential distribution with a mean of 120 (equal to the mean frequency of intake of orange juice/year from the LCPS) and rounded to the nearest integer for a total of 1,000 observations.

Simulation of predefined responses. To generate the predefined responses, we classified intake of orange juice into the Willett FFQ (2) categories: \(<12 \) times per year (once/month), \(12–36 \) times per year (2–3 times/month), \(52 \) times per year (once/week), \(104–208 \) times per year (2–4 times/week), \(260–312 \) times per year (5–6 times/week), \(365 \) times per year (once/day), and \(730 \) times per year (≥2 times/day). We then calculated the proportion of the 1,664 participants in each category. We used these proportions to create a multinomial distribution in which the probabilities for each predefined response category were equivalent to the proportion of the simulated open-ended responses for each category. Further details regarding development of the multinomial distribution of predefined responses are provided in Appendix 1.

Case and control simulations. To investigate the effect of using different predefined responses on the odds ratio for diet-disease associations, we simulated 200 replicates of 1,000 cases and 1,000 controls. First, we simulated open-ended responses from the same exponential distribution for orange juice as described above (mean = 120). We also simulated age as a normally distributed covariate (mean = 55 years; standard deviation, 12, rounded to the nearest integer) based on estimates from the LCPS control population. We then generated case or control status under 4 logistic regression models (Table 1) for different levels of orange juice intake (scaled to times/month), including age as a covariate. Note that, for model 1, the odds ratio referred to the percent decrease in risk from an increase in frequency of intake of orange juice, whereas, in model 2, we generated an increased risk effect. For model 3, the odds ratio referred to the decrease in risk after intake exceeded 78 times per year. We chose 78 times per year as a threshold because it is the midpoint of the gap between 1 time per week and 2–4 times per week. Model 4 applied the odds ratio to the same threshold of 78 times per year, but the odds ratio represented an increased risk effect. Further details regarding this simulation procedure are given in Appendix 2.

We applied a similar approach to simulate disease status based on the frequency of hamburger intake. We used an exponential distribution with a mean of 62 (equal to the mean frequency of hamburger intake per year from the LCPS).

Prior to conducting analyses, for the individuals whose frequencies of intake fell in a gap between predefined response categories, we used 1 of the following 7 methods to categorize that individual based on his or her frequency of intake: 1) the midpoint of the gap: those individuals in the gap whose frequency of intake was at or above the midpoint were included in the next higher category, and those whose frequency of intake was below the midpoint were included in the lower category; 2) minimum of the gap: we placed all individuals at or above the lowest intake in the gap into the

Table 1. Simulation Parameters for Disease Status

<table>
<thead>
<tr>
<th>Model</th>
<th>Intercept</th>
<th>Intake Effect* (Odds Ratio)</th>
<th>Age Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>-2.44</td>
<td>-0.22314 (0.80)</td>
<td>0.01</td>
</tr>
<tr>
<td>Model 2</td>
<td>-2.44</td>
<td>0.22314 (1.25)</td>
<td>0.01</td>
</tr>
<tr>
<td>Model 3</td>
<td>-2.44</td>
<td>-0.22314 ( I_{(intake &gt; 78)} ) (0.80)</td>
<td>0.01</td>
</tr>
<tr>
<td>Model 4</td>
<td>-2.44</td>
<td>0.22314 ( I_{(intake &gt; 78)} ) (1.25)</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\( I_{(intake > 78)} \) denotes an indicator function, taking the value 1 if intake is >78 and 0 otherwise.
Phase 2: pilot study

At The University of Texas M. D. Anderson Cancer Center, we piloted a study to develop a new core FFQ. Currently, we use an FFQ with open-ended responses, and, to facilitate recording responses, we attempted to modify the FFQ to include predefined categories for easier coding. The pilot study compared responses to the current version of the questionnaire—the current meat cookery FFQ—with the new version—the modified meat cookery FFQ. The current meat cookery FFQ recorded frequency of intake by using open-ended responses for various meats, whereas the modified meat cookery FFQ used predefined responses for frequency of intake of the same meats. For this study, we focused on intake of hamburger.

**Subjects.** Twenty-six participants completed the pilot study, a randomized, crossover design with 2 groups: group 1 received the current meat cookery FFQ at baseline (March 2008) and 6–8 weeks later (May 2008) received the modified meat cookery FFQ; group 2 received the modified meat cookery FFQ at baseline followed by the current meat cookery FFQ 6–8 weeks later. The pilot study sample was a convenience sample of M. D. Anderson Cancer Center employees. The study protocol was approved by The University of Texas M. D. Anderson Institutional Review Board.

**Statistical analysis of the pilot data.** We analyzed hamburger intake from the pilot data by treating the categorized intake as a scale ranging from 1 to 7, a ranking method similar to that in other dietary analyses. We computed a paired t test to compare the mean hamburger intake frequency estimated by the current meat cookery FFQ with that estimated by the modified meat cookery FFQ. P < 0.05 was considered significant.

### RESULTS

**Phase 1: simulation results**

**Comparing mean intake.** Table 2 shows type 1 error rates when we compared mean intake using open-ended responses with that obtained using predefined responses regarding frequency of drinking orange juice. We observed inflation in the type 1 error rate when the open-ended responses were categorized using any method different from that used to define the equivalent proportions. For example, for the n = 200 per group, the maximum of the gap method had the lowest type 1 error rate (25%), whereas the midrange, the midpoint, and the median of the methods had 66%, 79.5%, and 85% type 1 error rates, respectively. We observed the highest error rates in the midrange 1 of the gap method, at 97%, and the minimum of the gap method, at 100%. We observed similar trends for n = 1,000 and n = 40. In addition, as the sample size increased, so did the type 1 error rates, regardless of the method of categorizing the open-ended responses. However, when we defined the responses that fell into the gaps as “missing,” we observed acceptable type 1 error rates close to the nominal 0.05.

### Table 2. Type 1 Error Rate (Proportion of 200) Comparing Open-ended and Predefined Responses With Different Sample Sizes Using Simulated Orange Juice Intake

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Midpoint</th>
<th>Median</th>
<th>Quartile 1</th>
<th>Quartile 3</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Gap as Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.86</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.795</td>
<td>0.85</td>
<td>0.97</td>
<td>0.66</td>
<td>1.0</td>
<td>0.26</td>
<td>0.07</td>
</tr>
<tr>
<td>40</td>
<td>0.21</td>
<td>0.25</td>
<td>0.34</td>
<td>0.17</td>
<td>0.725</td>
<td>0.085</td>
<td>0.06</td>
</tr>
</tbody>
</table>
observed similar results for simulated hamburger intake (data not shown).

Estimating odds ratios. Table 3 shows the averaged odds ratios and 95% confidence intervals for orange juice for the 200 replicates generated from each of the 4 models in Table 1. We reported the odds ratios estimated by analyzing the original open-ended data on a times-per-month scale and then those estimated by analyzing the categorized open-ended data using the 6 methods shown in Figure 1 and the gap as missing method. For all models and all categorization methods, the 95% confidence interval for the average odds ratio excluded 1. For model 1, where the true odds ratio was 0.80, analyzing the data under its original scale of times per month was the only method that accurately estimated the true odds ratio (odds ratio (OR) = 0.8, 95% confidence interval (CI): 0.78, 0.82). Using the different methods to categorize observations that fell in the gaps into the predefined categories yielded varying odds ratios. Using the midpoint of the gap method (OR = 0.39, 95% CI: 0.36, 0.43) generated odds ratios similar to those using the median of the gap (OR = 0.42, 95% CI: 0.38, 0.46), quartile 3 of the gap (OR = 0.39, 95% CI: 0.36, 0.43), maximum of the gap (OR = 0.35, 95% CI: 0.31, 0.39), or gap as missing (OR = 0.36, 95% CI: 0.32, 0.41) methods. However, the odds ratio from the midpoint method was different from the odds ratio estimated by using the quartile 1 of the gap (OR = 0.45, 95% CI: 0.41, 0.48) or minimum of the gap (OR = 0.48, 95% CI: 0.44, 0.51) to categorize observations that fell in the gaps.

Note that the estimated odds ratios based on the categorized open-ended responses overestimated the magnitude of the true odds ratio, implying a much stronger protective effect than was simulated. Data analysis from model 2 showed similar trends, but the odds ratios were in the direction of risk, such that we observed elevated risk estimates. For the threshold models (models 3 and 4), none of the different categorizations of the open-ended data correctly estimated the true odds ratio (Table 3).

We also analyzed data simulated from orange juice intake with models 3 and 4, using the first category as a reference. For model 3, when we compared all categories of intake with the first level (less than once/month), the average odds ratios were equivalent across the midpoint, median, quartile 1, and quartile 3 categorizations (Table 4). For those categories below the threshold for disease risk (78 times/year), the 95% confidence intervals for the odds ratios included 1. For those categories above the threshold, the methods recovered the true simulated odds ratio. However, when we used the “minimum of the gap” method, the average odds ratio was significantly different from the odds ratios obtained from the other methods in the 2–4 times per week category (P < 0.01). When we used the “maximum of the gap” method, the average odds ratio for the once per week category was significantly different from the odds ratios obtained when other methods were used to allocate the observations in the gaps (P < 0.01). This result shows that, when categorizing the data, if the true threshold was within the category, then the estimation of the odds ratio was biased. Similar results were found with the average odds ratios for model 4 (Table 5), with the odds ratio being in the direction of inferring risk.
Phase 2: pilot study results

The average intake of hamburgers per year as recorded by the current meat cookery FFQ was 62.23 (standard deviation, 67.56), which fell in the gap between categories 3 and 4 (Figure 1). The median intake of hamburgers as reported on the current meat cookery FFQ was 43.5, which fell in the gap between intake categories 2 and 3 (Figure 1). The average and median intake of hamburgers according to the modified meat cookery FFQ (treating the categories as intervals) was 2.0 (standard deviation, 0.75). We used each of the 7 categorizations for the open-ended responses about hamburger intake collected with the current meat cookery FFQ for the pilot study. Table 6 summarizes mean hamburger intake recorded by the current meat cookery FFQ categorized by each of the methods. For all categorization methods, except the one defining responses of individuals who fell into the gaps as “missing,” we found a significant difference between the mean intake calculated from the current meat cookery FFQ and that calculated from the modified meat cookery FFQ. The mean of the current meat cookery FFQ varied according to the method used to categorize the observations of the open-ended data in the gaps. Table 6 also shows the results for the paired t test comparing the mean hamburger intake using the current meat cookery FFQ versus the modified meat cookery FFQ, with responses on the current meat cookery FFQ categorized by the different methods. The variation across categorization methods implied that, even with actual data, the method chosen to categorize the data influences the inference about mean differences.

**DISCUSSION**

The simulations showed that the method used to categorize the open-ended response strongly affected type I error when we compared mean intake between open-ended and predefined responses. We also observed that the methods chosen to categorize open-ended responses biased odds ratio estimation, whether the data were treated as purely categorical or as interval. When we compared intake reported by open-ended responses versus predefined responses, we noticed a variation in the P values reported, and this variation depended on how we categorized the open-ended responses into the noncontiguous, predefined responses.

By simulating the open-ended and predefined responses to have the same proportions, we observed that the method of categorizing continuous observations that fall between predefined categories inflates type I error. Only one method for categorizing the data did not severely inflate type I error: treating the data in the gaps as missing. We caution against concluding that this approach is the best for comparing means; in our analysis, this method precisely matched the method we used to simulate the data. As shown in Appendix 2, our use of proportions in the multinomial distribution matched the proportion of the open-ended data that fell into each of the predefined categories, essentially ignoring those observations in the gap. These results imply that, when making inferences about the mean difference and
comparing categorized open-ended data with predefined categorical data, one needs to categorize open-ended data by using a method identical to the thought process a participant would use when categorizing his or her intake into 1 of 2 noncontiguous, predefined categories. However, this categorization could be accomplished only by conducting cognitive interviews with participants.

Our simulation studies of diet-disease associations using logistic regression models to estimate the odds ratios showed that categorizing open-ended data by using any method biased the estimates away from the null. This result is unsurprising, because categorizing the data can be seen as changing the size of the unit change in intake, which corresponds to a larger odds ratio (17). However, once the open-ended responses are categorized, the method of categorizing the open-ended responses that fall in a gap does not have as strong an effect on odds ratios as it does on inferences about differences in mean intake. The odds ratio estimates for models 1 and 2 showed some variation, but, given the 95% confidence intervals of each estimate, the inference regarding the significance of the odds ratios was similar across categorization methods.

When the true underlying mechanism of risk is a threshold, as simulated in models 3 and 4, treating the categories as ordinal does not detect any change in risk due to intake. Only by comparing each category with a reference can the analysis correctly estimate the simulated odds ratio; however, the odds ratio estimate differs depending on the method of categorizing the data. Most categorization methods tell us that the threshold for orange juice intake to affect risk is between once per week and 2–4 times per week, but the maximum and minimum methods do not make this finding evident. Recall that the true threshold is the midpoint of the gap between once per week and 2–4 times per week. The minimum and maximum place the entire gap into one of the adjacent categories. Therefore, logistic regression estimates an odds ratio for a group that consists of individuals on both sides of the true risk threshold, thereby biasing the estimate toward the null (3, 5, 7).

Given the results from the simulated data, interpretation of the results comparing mean intake collected from an FFQ using open-ended responses with mean intake collected

### Table 5. Average Odds Ratios for Categorical Analysis of Categorized Open-ended Data Simulated From Model 4 Using Simulated Orange Juice Intake

<table>
<thead>
<tr>
<th>Category</th>
<th>Midpoint</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Midpoint</th>
<th>Median</th>
<th>Quartile 1</th>
<th>Quartile 3</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current Meat Cookery FFQ</td>
<td>2.58 (1.14)</td>
<td>2.69 (1.23)</td>
<td>2.81 (1.20)</td>
<td>2.58 (1.14)</td>
<td>3.00 (1.47)</td>
<td>2.38 (1.10)</td>
<td>2.40 (1.40)</td>
<td>2.46 (1.35)</td>
<td>2.40 (1.35)</td>
<td>2.35 (1.35)</td>
</tr>
</tbody>
</table>

### Table 6. Results for Paired t Test for Pilot Study of Hamburger Intake Under Different Methods of Categorizing Gaps

<table>
<thead>
<tr>
<th>Method for Gap</th>
<th>No.</th>
<th>Current Meat Cookery FFQ Mean (SD)</th>
<th>Current Meat Cookery FFQ-Modified Meat Cookery FFQ</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midpoint</td>
<td>26</td>
<td>2.58 (1.14)</td>
<td>0.59</td>
<td>0.0021</td>
</tr>
<tr>
<td>Median</td>
<td>26</td>
<td>2.69 (1.23)</td>
<td>0.69</td>
<td>0.0005</td>
</tr>
<tr>
<td>Quartile 1</td>
<td>26</td>
<td>2.81 (1.20)</td>
<td>0.81</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Quartile 3</td>
<td>26</td>
<td>2.58 (1.14)</td>
<td>0.58</td>
<td>0.0021</td>
</tr>
<tr>
<td>Minimum</td>
<td>26</td>
<td>3.00 (1.47)</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Maximum</td>
<td>26</td>
<td>2.38 (1.10)</td>
<td>0.38</td>
<td>0.0221</td>
</tr>
<tr>
<td>Missing</td>
<td>14</td>
<td>2.46 (1.35)</td>
<td>0.40</td>
<td>0.0558</td>
</tr>
</tbody>
</table>

Table 6. Results for Paired t Test for Pilot Study of Hamburger Intake Under Different Methods of Categorizing Gaps

Abbreviations: FFQ, food frequency questionnaire; SD, standard deviation.
from an FFQ using predefined responses for hamburger intake is not straightforward. Comparing mean intakes by using simulated data showed an inflation of type 1 error. Therefore, the significant difference detected by the t test comparing the current meat cookery FFQ with the modified meat cookery FFQ may be artificially generated by the method chosen to categorize the open-ended data before the comparison was made.

A limitation of this research is that we could not address the participant’s framework for choosing answers or the error introduced by high quantities of missing data as a result of mixing open-ended responses after a series of closed-ended questions, because the simulations model data that had already been collected. By modeling our simulations as we did, we isolated the error induced by the researcher’s choice of how to pool the data and showed that treatment of gaps can modify the inference in a study independently of more common sources of variation in a nutrition epidemiology study using FFQs.

In summary, when mean intakes based on open-ended responses are compared with those measured by predefined categories, the type 1 error rate can be severely inflated. Therefore, significant differences in mean intakes between 2 groups collected by using these 2 different methods cannot be trusted. In contrast, when odds ratios are estimated, if the risk changes continuously with intake, the odds ratio estimates are similar regardless of the categorization methods used, avoiding the extremes of the maximum or minimum methods. Yet, when there is an intake threshold for the change in risk, then if the threshold occurs within a large category, the precise magnitude of the threshold may be masked by the method selected to categorize the gaps.

Therefore, the message is to pool data of the same response types when possible. If different response types are used, odds ratios will have a slight, but most likely acceptable, bias. Mean intake cannot be compared when pooling data. Therefore, researchers must proceed with caution when building risk models. Commonly, t tests for mean intake are computed first to determine whether a dietary variable enters multivariable logistic regression models. When dealing with pooled data of different types, researchers must rely on other multivariable model-building methods based on odds ratios such as stepwise regression or Bayesian model averaging (18) rather than the preliminary t tests. When developing the questionnaire, one can minimize the effect of the error from categorizing the gaps by simply wording the predefined categories so they are contiguous.

ACKNOWLEDGMENTS

Author affiliation: Department of Epidemiology, University of Texas, M. D. Anderson Cancer Center, Houston, Texas (Michael D. Swartz, Michele R. Forman, Somdat Mahabir, Carol J. Etzel).

This work was supported in part by National Cancer Institute grant K07CA123109 to M. D. S. and grants K07CA093592 and CA123208 to C. J. E.

The authors thank the Lung Cancer Study research team, who collected the data, as well as the Nutrition Epidemiology working group, who provided the nutrition data on which the simulations were based. They thank Dr. Margaret R. Spitz for allowing access to the data, collected under National Cancer Institute grant CA55769.

Conflict of interest: none declared.

REFERENCES

18. Swartz MD, Yu RK, Shete S. Finding factors influencing risk: comparing Bayesian stochastic search and standard variable

**APPENDIX 1**

**Simulating Predefined Responses From a Multinomial Distribution**

This Appendix provides additional details for simulating 2 groups with equal intake, one reporting intake with open-ended responses and the other with predefined responses. As mentioned in the paper, we simulated 1,000 open-ended responses from an exponential distribution with mean = 120 and rounded to the nearest integer. We matched the exponential distribution with mean = 120 to a categorical multinomial distribution using a 3-step process. First, we used the same exponential distribution (mean = 120) to generate 2 million responses (to approximate the true distribution) regarding frequencies of orange juice intake per year. Second, we estimated the proportion of frequencies in each of the following categories shown in Figure 1: less than 12 times per year (once/month), 12–36 times per year (2–3 times/month), 52 per year (once/week), 104–208 times per year (2–4 times/week), 260–312 times per year (5–6 times/week), 365 times per year (once/day), and 730 times per year or more (≥2 times/day). Third, we computed the 1,000 predefined responses from a multinomial distribution, with the probabilities for each category being equal to the proportion estimated from the exponential distribution. We repeated the same procedure with mean = 62 to simulate hamburger intake.

**APPENDIX 2**

**Simulating Cases and Controls**

We simulated a population of 20,000 individuals using the distributions described in the Materials and Methods section of the text for age (normal distribution, with mean = 55 (standard deviation, 12), rounded to the nearest integer) and orange juice intake (exponential distribution with mean = 120, rounded to the nearest integer). We scaled intake to times per month (times/year divided by 12) and calculated the probability of disease for each of the simulated covariate pairs of age and intake using the logistic models from Table 1. We then simulated disease status from these probabilities. We then sampled 1,000 cases from those with the simulated disease and 1,000 controls from those without the simulated disease. We repeated this procedure with a mean of the exponential distribution = 62 to simulate hamburger intake.