Published in final edited form as:

Stat Med. 2008 August 30; 27(19): 3789–3804. doi:10.1002/sim.3281.

Modeling heaping in self-reported cigarette counts

Hao Wang* and Daniel F. Heitjan

Department of Biostatistics & Epidemiology, University of Pennsylvania, Philadelphia, PA 19104, U. S. A.

SUMMARY

In studies of smoking behavior, some subjects report exact cigarette counts, whereas others report rounded-off counts, particularly multiples of 20, 10 or 5. This form of data reporting error, known as *heaping*, can bias the estimation of parameters of interest such as mean cigarette consumption. We present a model to describe heaped count data from a randomized trial of bupropion treatment for smoking cessation. The model posits that the reported cigarette count is a deterministic function of an underlying precise cigarette count variable and a heaping behavior variable, both of which are at best partially observed. To account for an excess of zeros, as would likely occur in a smoking cessation study where some subjects successfully quit, we model the underlying count variable with zero-inflated count distributions. We study the sensitivity of the inference on smoking cessation by fitting various models that either do or do not account for heaping and zero inflation, comparing the models by means of Bayes factors. Our results suggest that sufficiently rich models for both the underlying distribution and the heaping behavior are indispensable to obtaining a good fit with heaped smoking data. The analyses moreover reveal that bupropion has a significant effect on the fraction abstinent, but not on mean cigarette consumption among the non-abstinent.

Keywords

Bayesian inference; heaped data; rounded data; smoking cessation; zero-inflated negative binomial; zero-inflated Poisson

1. Introduction

This paper considers the analysis of data on reported cigarette consumption from a recent clinical trial evaluating the effect of the antidepressant drug bupropion in promoting smoking abstinence [1,2,3]. The study randomized 555 regular smokers to receive behavioral counseling plus 10 weeks of either twice-daily bupropion or placebo. All subjects were to quit smoking on the target quit date (TQD), two weeks after the start of medication. Telephone assessments followed at end of treatment (EOT) (eight weeks post-TQD), and at 6 and 12 months post-TQD. At each follow-up assessment, participants reported the number of cigarettes smoked each day since the previous assessment using a time-line follow-back (TLFB) method [4].

Self-reported cigarette consumption data are typically *heaped*, in the sense that the cigarette counts are evidently recorded rounded off to different levels of precision [4,5,6]. Although heaping is well known to be a troubling nuisance in such data, its magnitude is difficult to quantify and its effects on statistical inferences are poorly understood.

*Correspondence to: Department of Biostatistics & Epidemiology, University of Pennsylvania, School of Medicine, Philadelphia, PA 19104, U. S. A..

To illustrate the extent of the problem, we present in Figure 1 histograms of cigarette counts at day 56 (EOT) from the two arms of the bupropion trial. In addition to the large number of zeros (potentially representing subjects who quit smoking), there are evidently heaps at multiples of five cigarettes, most noticeably at 20. Heaping is likely to arise as a particular form of recall error in the retrospective TLFB method; that is, subjects may not remember precisely how many cigarettes they smoked each day, and therefore report a rounded approximation. On the other hand, the fact that cigarettes in the United States can be purchased only in packs of 20 may promote the practice of smoking a pack or a half-pack per day, in which case the heaps at 20 and 10 may represent actual consumption values. In any event, current approaches to the analysis of cigarette count data take the data at face value despite the clear departures of such histograms from the shapes seen in standard count-data models.

Heaping is of concern because it distorts the distribution of the observed counts and thus may bias estimation. To demonstrate this point, we evaluated through a putative model the impact of heaping on mean cigarette count. We calculated the mean for heaped counts that arise as observations from an underlying Poisson distribution under the assumption that each observation is rounded to the nearest multiple of 20 with a given probability. The difference between the mean of the heaped counts and the mean of the underlying Poisson distribution is thus the bias from heaping. From Table I, we see that the heaping bias depends on both the true Poisson mean and the fraction of observations that are heaped. Moreover, the potential effect of heaping in estimating mean cigarette counts is substantial. For example, when 50% of the data are heaped, the bias from heaping is 1.32 for a true mean of 15, but -0.68 for a true mean of 25. Thus, if one were to use 50% heaped data to compare two treatment arms, whose true means are 25 and 15, the treatment effect estimate would be attenuated by 20%.

Heaping occurs in a number of applied contexts, particularly those involving retrospective collection of self-reported data items. Some examples include age [7], gestational age [12, 13], age at menopause [17], breastfeeding duration [8], household total expenditure [9], unemployment duration [10], length of emergency room visits [11], blood pressure [14,15], and numbers of drug or sexual partners [16].

The literature on heaping sprawls over statistics and a range of application areas, with relatively little work to date on formal modeling [18,19,12,20,21]. One prominent stream of research seeks to test whether the evidence for heaping in a dataset is statistically significant [16], whereas other approaches seek to quantify the extent of heaping or to smooth out the heaps [22,23].

The first attempt to rigorously model heaping was by Heitjan and Rubin [18], who studied the effect of heaped age reporting on inferences about the nutritional status of Tanzanian children. In their data, only a few of the youngest children had ages reported exactly (i.e., to the nearest month); most reported ages greater than 12 months fell on full or half years. The authors hypothesized that the population consisted of different types of age rounders (exact, nearest half-year or nearest full-year), and that the probability of coarse rounding increased with age. They estimated a range of models incorporating these assumptions and imputed multiple sets of data under each. They concluded that assumptions about the rounding could substantially affect inferences, although simple imputation schemes worked about as well as more complicated model-based methods. An article by Pickering [12] modeled the probability of delivery of a child as a function of its gestational age, whose reporting was subject to heaping at even numbers of weeks. The model consisted of a distribution for misclassification of odd as even ages, superimposed through a composite link function on a GLM for the outcome. Pickering concluded that although the heaping had a substantial effect on model fit and predicted event counts, it did not affect the estimated relative risks of important predictors.

Later, Dellaportas [20] and Wright and Bray [21] proposed a mixture model with components corresponding to rounding and exact measurement for applications in perinatal medicine.

More recently, Farrell *et al.* [24] modeled heaped cigarette counts in an English health survey. Arguing from microeconomic principles, they proposed that addicted smokers seek to regulate consumption through their purchasing habits, making it plausible that individuals will smoke a pack (10 or 20 cigarettes in the UK) or a half-pack (5 or 10 cigarettes) per day, the patterns commonly observed in the survey. They then partitioned smoking behavior into several classes: non-smokers, free smokers (those who smoke and report a variable number of cigarettes per day), and groups who are “attracted” to one of the common cigarette consumption heaps — 5, 10, 15, 20 or more than 20 cigarettes per day. Their model involved mixture of a Poisson distribution with a multinomial process indicating smoking behavior classes. They parameterized the model with a logistic regression for the class identity and a Poisson regression for the mean. By contrast, the models of Heitjan and Rubin [18], Pickering [12], Dellaportas *et al.* [20] and Wright and Bray [21] describe the heaps as the result of a process of rounding to nearby popular values. Moreover, Heitjan and Rubin [18] and Pickering [12] applied both ignorable (outcome is unrelated to rounding probability) and nonignorable (rounding probability depends on outcome) versions of their models.

Our aims in this article are threefold. First, we present a comprehensive modeling strategy for heaped data. Our approach has three elements: (i) The true latent outcome with its distribution, (ii) the rounding behavior, whose distribution possibly depends on the value of the latent outcome according to a selection model, and (iii) a mapping from the latent outcome and rounding variate to an observed (heaped) value. The intention is to use the full model fit to extract valid estimates of the parameters of the distribution of the latent outcome. The dependence of the heaping probability on the underlying cigarette count (a feature that is absent from some recently proposed models) is potentially important because it captures a critical aspect of the data — i.e., that people who smoke many cigarettes are less likely to remember precise counts. Second, we implement the modeling in a Bayesian framework, the first heaping selection model to be so fit. Among its other benefits, this allows us to compare the first of both nested and non-nested pairs of models. Third, our model for the latent cigarette counts is of particular scientific interest because it resolves the treatment effect into effects on the probability of abstinence and the mean count given non-abstinence, yielding a clearer picture of the treatment’s potential clinical usefulness.

2. Methods

2.1. Models of heaped data

Let X denote the latent outcome variate and G the rounding behavior. With smoking, we might assume that G takes four possible values: $G = 1$ for exact reporting, $G = 2$ for rounding to the nearest multiple of 5, $G = 3$ for rounding to the nearest 10, and $G = 4$ for rounding to the nearest 20. The model asserts that Y , the reported cigarette count at EOT, is a function of X and G :

$$Y = Y(X, G).$$

For example, a subject with $X = 14$ and $G = 1$ reports $Y(14, 1) = 14$, whereas $Y(14, 2) = 15$, $Y(14, 3) = 10$, and $Y(14, 4) = 20$. Figure 2(a) illustrates this heaping mechanism.

Any coarsened outcome y may arise from one or more possible (x, g) pairs. We denote the set of such pairs as $XG(y) = \{(x, g) : y = Y(x, g)\}$. For example, a reported consumption of $y = 5$ may represent a precise unrounded value $((x, g) = (5, 1))$ or rounding across a range of nearby values $((x, g) \in \{(3, 2), (4, 2), (5, 2), (6, 2), (7, 2)\})$. The likelihood function for an observed

y is the sum of the probabilities of the (x, g) pairs that would give rise to it [25,26,27]. If we designate the model for X as $f_X(x; \theta)$ and the model for G given $X = x$ as $f_{G|X}(g/x; \gamma)$, where θ is the parameter of the X model and γ is that of the G/X model, the likelihood is

$$L(\theta, \gamma; y) = \sum_{(x, g) \in XG(y)} f_X(x; \theta) f_{G|X}(g/x; \gamma).$$

Next we propose specific forms for the model components $f_X(x; \theta)$ and $f_{G|X}(g/x; \gamma)$.

ZIP and ZINB models for cigarette count—The Poisson distribution is a natural starting point for modeling count data, and to accommodate potential overdispersion we also consider the negative binomial distribution. A salient characteristic of cigarette data from smoking cessation studies, as illustrated in Figure 1, is the tendency toward excess zeros. Zeros in this context may result from either rounding of small cigarette counts, occasional non-smoking days, or complete abstinence. Excess zeros representing abstinence, should they exist, are a potentially important object of modeling interest. To accommodate it, we have adopted zero-inflated (ZI) versions of our count-data models [28,29]. The ZI model describes the data as a mixture of a full count distribution with a point mass at zero, allowing us to simultaneously estimate treatment effects on the fraction of abstainers and the rate of smoking among non-abstainers.

Combining these elements, we define our zero-inflated Poisson (ZIP) model for the true underlying cigarette count X as

$$f_X(x; \theta) = (1 - \theta_1)I_0(x) + \theta_1 \exp(-\theta_2) \theta_2^x / x!.$$

where $I_0(x)$ is an indicator that $x = 0$. That is, we assume that a fraction θ_1 of subjects continues to smoke with mean cigarette count θ_2 , while a $1 - \theta_1$ fraction abstains. We can model θ_1 and θ_2 as functions of treatment and other predictors:

$$\text{logit}(\theta_1) = V\eta, \quad \log(\theta_2) = W\beta, \quad (1)$$

where V is a matrix of predictors of the fraction non-abstinent and W is a matrix of predictors of mean cigarette count among the non-abstinent. The corresponding zero-inflated negative binomial distribution (ZINB) follows the form

$$f_X(x; \theta) = (1 - \theta_1)I_0(x) + \theta_1 \frac{\Gamma(x + \frac{1}{\theta_3})}{\Gamma(x+1)\Gamma(\frac{1}{\theta_3})} (1 + \theta_2\theta_3)^{-\frac{1}{\theta_3}} \left(1 + \frac{1}{\theta_2\theta_3}\right)^{-x},$$

where $\theta_3 \geq 0$ is a dispersion parameter. Here again, θ_1 represents the proportion of non-abstainers, whose mean cigarette count is θ_2 . The variance of X given non-abstinence is $\theta_2(1 + \theta_2\theta_3)$, and consequently if $\theta_3 > 0$ the variance exceeds the mean. The NB model arises as a Gamma mixture of Poissons, where θ_3 is the variance of the Gamma distribution describing the heterogeneity and the ZINB density converges to the ZIP as $\theta_3 \rightarrow 0$. We can model the fraction of non-abstainers and the mean given non-abstinence exactly as in (1). We assume for convenience that the overdispersion parameter θ_3 does not depend on predictors.

A model for heaping behavior—In the Heitjan-Rubin [18] age heaping model, the assumed latent age and heaping behavior followed a bivariate normal distribution conditional

on covariates, with ranges of g corresponding to designated rounding behaviors. We seek here to estimate similar models using a refinement of the Heitjan-Rubin approach. From the histogram of reported cigarette counts (Figure 1), there are evidently extra observations (i.e., heaps) not only at 20, 40 and 60 (representing one, two and three packs, respectively), but also at 10, 30 and 50, and to a lesser extent at 5, 15, 25 and 45. Although there are some subjects at every count of 20 or less, there is an apparent attraction to reporting multiples of five, with multiples of ten potentially more attractive and multiples of 20 most attractive of all. Thus we propose a model in which the heaping behavior variable takes on four possible values: $g = 1$ for reporting the exact count, $g = 2$ for rounding to the nearest multiple of 5, $g = 3$ for rounding to the nearest multiple of 10, and finally $g = 4$ for rounding to the nearest multiple of 20.

The age data of Heitjan and Rubin [18] showed a trend toward increased coarseness with increasing X , and Figure 1 suggests that this may occur with cigarettes as well. To describe this behavior, we assume that reporting behavior depends on X according to a proportional odds model:

$$f_{g|x}(g|x;\gamma) = \begin{cases} (1+\exp(\gamma_1+\gamma_0 x))^{-1}, & \text{if } g = 1; \\ (1+\exp(\gamma_2+\gamma_0 x))^{-1} - (1+\exp(\gamma_1+\gamma_0 x))^{-1}, & \text{if } g = 2; \\ (1+\exp(\gamma_3+\gamma_0 x))^{-1} - (1+\exp(\gamma_2+\gamma_0 x))^{-1}, & \text{if } g = 3; \\ 1 - (1+\exp(\gamma_3+\gamma_0 x))^{-1}, & \text{if } g = 4. \end{cases}$$

The model posits a separate logistic regression on x for each cut point in g , with common slope γ_0 and intercepts γ_j ; $j = 1, 2, 3$ for cuts at $g = 1, 2, 3$, respectively, where the intercepts are constrained to satisfy $\gamma_1 > \gamma_2 > \gamma_3$. A value of $\gamma_0 = 0$ implies that the rounding rate does not depend on the underlying value of x .

Summing up, in the proposed model $\theta = (\theta_1, \theta_2)$ is the parameter of interest and $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ (or $\gamma = (\gamma_3, \gamma_0, \gamma_1, \gamma_2, \gamma_3)$ with the ZINB model) is a nuisance parameter.

2.2. Bayesian inference for the heaping model

We estimate the model by a Bayesian approach involving numerical maximization of the posterior followed by importance sampling. We use proper priors that place mass in likely parts of the parameter space but are sufficiently diffuse to allow the possibility that the truth lies elsewhere. The use of proper priors also obviates the need to verify the existence of the posterior. For convenience, we assume that all the parameters are *a priori* independent.

Priors for the ZINB and ZIP models—We use logistic regression to model the proportion abstinent in both ZIP and ZINB models, with the two parameters in the logit model representing the log odds of non-abstinence in the placebo group (β_1) and the log odds ratio of non-abstinence on bupropion versus placebo (β_2). Previous studies suggest that the probability of abstinence at EOT with no medical treatment is about 10% (or less), while that for bupropion is near 30% [30,31,32]. Because the logistic regression parameters lie in the range $(-\infty, \infty)$, it is natural to adopt normal priors. Here we take the prior mean of β_1 to be 2.2 (corresponding to 90% non-abstainers on placebo) and the mean of β_2 to be -1.35 (corresponding to 3.86 times higher odds of quitting on bupropion than placebo), both with large variance 10^2 . Similarly, assuming that average cigarette consumption is 10 for those who do not quit, we choose the priors $N(2.3, 10^2)$ for the log mean count in the placebo group (η_1), and $N(0, 10^2)$ for the log ratio of means on bupropion versus placebo (η_2). The ZINB model has the further parameter θ_3 measuring overdispersion. Recognizing that ZINB allows the variance to exceed the mean θ_2 by $\theta_3\theta_2^2$, we select an exponential prior with mean 0.1 for θ_3 . Given mean cigarette count

θ_2 , the average variance would be $\theta_2(1 + 0.1\theta_2)$. If the mean cigarette consumption were 10, the NB variance would thus double the Poisson variance.

Prior for heaping behavior models—We have little prior data on heaping in cigarette counts, so we are less certain about the likely values of the parameters $\gamma_j, j = 0, 1, 2, 3$, for the heaping portion of the model. We therefore set each of them to be $N(0, 10^2)$, subject to the constraint that $\gamma_1 > \gamma_2 > \gamma_3$.

Posterior computation—We use importance sampling [33,34] to simulate from the posterior distribution of the parameters. The steps are as follows: We first compute the posterior mode and information using a Nelder-Mead simplex algorithm [35] with finite-difference derivatives [36]. We then estimate the posterior with a multivariate t_5 density whose mean equals the posterior mode and whose dispersion equals the inverse of the posterior information matrix at the mode. We then draw a large number (4,000) of samples from this multivariate t_5 proposal distribution, at each draw computing the ratio w of the true posterior density to the proposal density, the so-called importance ratio. We evaluate posterior moments by averaging functions of the simulated parameter draws with the importance ratios w as weights.

We compare model fits using Bayes factors [37]. The Bayes factor (BF) for comparing models M_1 and M_2 is the ratio of the marginal probability of the observed data d under these models:

$$\text{BF} = \frac{p(d|M_2)}{p(d|M_1)}.$$

We also use importance sampling to compute the Bayes factors [38]. Under model M , $p(d|M)$ is the normalizing constant z for the posterior density $p(\theta|d)$. We denote $q(\theta)$ as the posterior density known up to the normalizing constant z (i.e. $p(\theta|d) = q(\theta, d)/z$), and $p^*(\theta)$ as the approximation to the true posterior $p(\theta|d)$. Here the density $p^*(\theta)$ is completely known. The importance sampling estimator of z is based on the identity

$$z = E_{p^*} \left[\frac{q(\theta)}{p^*(\theta)} \right]$$

and the Monte Carlo estimate of $p(d|M)$ is

$$\widehat{p}(d|M) = \widehat{z} = \frac{1}{m} \sum_{i=1}^m \frac{q(\theta^{(i)})}{p^*(\theta^{(i)})}$$

where $\theta^{(1)}, \dots, \theta^{(m)}$ are the draws from $p^*(\theta)$. Taking the multivariate t_5 proposal density as the importance sampling function $p^*(\theta)$, we can take advantage of the importance ratios w computed previously, calculating $\widehat{p}(d|M)$ as their mean.

Computer code—We have written functions in R (Version 2.3.1) to perform all the calculations. Code is available from the second author's web site, <http://www.cceb.upenn.edu/heitjan>.

3. Results

To evaluate the modeling approach and study the sensitivity of inferences, we fit a series of models including standard and ZI Poisson and NB models, and models ignoring and accounting

for heaping. In addition to the heaping behavior assumed in Figure 2(a), we consider an alternative, potentially more plausible mechanism that eliminates the behavior “rounding to 0” (Figure 2(b)). In all cases the targets of inference were the proportion abstinent and the mean cigarette consumption given non-abstinence

Table II lists BFs for pairwise comparison of the models. The evidence is strong that the models with heaping are preferable in all circumstances. In particular, the ZINB model with heaping fits much better than any other model considered, with twice the natural logarithm of BF equal to 1583 over the heaped ZIP model and 715 compared to the unheaped ZINB model. The BF comparing any one of the NB models with the heaped ZIP model indicates strong evidence in favor of NB, suggesting that both the appropriate underlying distribution and heaping model are needed to obtain a good fit. Comparing heaping behavior assumptions (1) and (2), twice the natural logarithm of BF is 5.7 from the ZINB models and -13.3 for the ZIP models; thus heaping model (1) is preferable under ZINB whereas model (2) is preferable under ZIP.

Bayes factors rely on prior distributions, and thus can be sensitive to the choice of prior. Consequently, we also calculated approximate BFs using the Schwarz criterion [37], which avoids the introduction of priors. Results (not shown) were similar.

Figure 3 shows plots of the residuals (observed minus expected counts on the square-root scale). Although the heaping artifact is reduced substantially in the heaped ZIP model, there is still considerable lack of fit apparent in it, suggesting inadequacy of the Poisson assumption. By contrast, the heaped ZINB model not only smooths out the heaps, it also largely corrects the lack of fit in the unheaped ZINB model. Although difficult to see from the graphs in Figure 3, there are small excesses at even-numbered counts in the residual plots from the heaped ZINB models, suggesting that there is some preference for even cigarette counts even beyond the evident heaping at multiples of 10 and 20.

Table III presents estimates of mean cigarette consumption (in the non-abstinent fraction under the ZI models; else overall) under the estimated models (see also Figure 4). Adding the ZI component increases the mean estimate substantially by eliminating the large proportion (30%–40%) of subjects who are identified as abstainers. Moving from the Poisson to the NB distribution has little effect on the posterior mode but substantially widens the 95% probability interval. In the ZINB models, incorporating heaping mechanism (1) decreases the estimates of mean consumption by 1.1 cigarettes for both placebo and bupropion arms, which corresponds to 8.7 and 8.3% differences in the posterior mode. Their 95% intervals overlap and the heaping model shows a slightly wider interval. The difference is more pronounced in the ZIP model, where the point estimate declines as much as 14%, and the interval of the mean adjusted for heaping shifts to the left and does not overlap with the interval from the unadjusted model. The ZINB model with the rounding mechanism (2) gives estimates close to the ZINB model without heaping.

We estimate the treatment effect on smoking rate by the ratio of mean cigarette counts (overall or in the non-abstinent fraction). The simple models produce much smaller ratios because they do not distinguish abstinent and non-abstinent counts. The rate ratio is the same in all other models, and the credible interval length is unaffected by the assumption of heaping. Thus the estimated treatment effect on smoking rate is robust to heaping model specification.

Modeling heaping has considerable impact on estimation of the quit probability. From Table IV, the non-abstinent fractions estimated by the models with heaping are larger than those of the models without (see also Figure 4). The non-abstinent proportion in the placebo arm from the heaped ZINB model is 8% higher than in the unheaped ZINB model (70% vs. 62%). The odds ratio for quitting is also sensitive, with heaping models showing as much as a 14%

difference compared to their non-heaping counterparts. Probability intervals largely overlap, however.

The posterior mode of the overdispersion parameter θ_3 is 0.508 (95% probability interval [0.435, 0.600]) from the unheaped ZINB model and 0.607 ([0.503, 0.739]) in the heaped ZINB model. Incorporating heaping in the model reveals greater heterogeneity in the data.

Although it is not of direct interest for the application, the logistic regression slope γ_0 is important in that it represents the potential association of heaping behavior with underlying cigarette consumption. In the heaped ZINB model, we estimate the posterior mode of γ_0 as 0.104, with 95% probability interval [0.060, 0.157]. The estimate from the heaped ZIP model is larger with a tighter 95% interval (0.118 [0.089, 0.151]). The positive slope implies that the chance of coarser heaping increases with the latent cigarette count, in line with our intuition and our impression from Figure 1.

Figure 5 presents the estimated heaping probabilities from ZINB heaping model (1) as a function of the true cigarette count. As suspected, the probability of rounding coarsely increases rapidly with true count. Averaging across the cigarette count distributions, the probabilities of exact reporting and rounding to the nearest multiple of 5, 10 and 20 are estimated as 0.52, 0.19, 0.15 and 0.14 on placebo, and 0.58, 0.18, 0.13 and 0.12 on bupropion. Thus the fraction of rounded subjects is 40% or more, suggesting that a failure to model heaping could substantially affect estimates of mean cigarette count.

To investigate sensitivity to the priors, we re-estimated the ZINB heaped (1) model under a range of prior distributions. For example, prior data suggested $N(-1.35, 10^2)$ as the prior for the parameter log odds ratio of non-abstinence on bupropion compared to placebo. A conservative prior assumption is that bupropion treatment is equally likely to increase or decrease abstinence, so we also tried $N(0, 10^2)$ as the prior for this parameter. We used $N(0, 10^2)$ for the parameter log odds of non-abstinence in the placebo group (where a mean of 0 corresponds to a 50% quit rate). For log mean cigarette consumption among non-abstainers, we tested $N(2.996, 10^2)$, which assumes that the mean count is centered at 30. We also used normal priors with considerably tighter distributions ($\sigma_N = 2$ and 0.6) on all the parameters including those describing heaping behavior. Besides the normal, we also tried the Cauchy, or $t_1(\mu, \sigma_c^2)$ distribution for parameters β and η . Specifically, we used the prior $t_1(0, 1:099^2)$ on the log odds ratio of non-abstinence, which assigns probability 0.5 to $\{1/3 < \text{OR} < 3\}$ [39]. We also tried a $t_1(0, 0.459^2)$ as the prior for the log odds of non-quitting in the placebo group; this assigns the same probability to $(\mu - 10, \mu + 10)$ as does $N(0, 10^2)$. With regard to the dispersion parameter θ_3 , we ran an exponential $\varepsilon(1)$ to allow a larger variance of X given non-abstinence.

The results of our sensitivity analysis (Table V) based on the ZINB heaped (1) model show that the posterior modes are reasonably stable under this range of priors, except for the tightest normal priors with $\sigma_N = 0.6$. Under this prior, the posterior mode of the OR of non-abstinence is 0.28 with the non-abstinent fraction 0.81 in the placebo arm, while the mean smoking rate among non-abstinent is only 10.3; the posterior mode of the rate ratio is more stable.

Lack of fit in the Poisson and ZIP models could be due to the failure to consider available predictors that might explain extra-Poisson variation. To assess this, we re-estimated the Poisson and NB models including as predictors the baseline values of the Fagerstrom Test of Nicotine Dependence (FTND) and the CES-D depression score. Including these parameters had only a modest effect on the others (results not shown).

Our approach assumes that the heaps constitute a mixture of genuine latent consumption and rounded recall errors, where we have used the ZIP and ZINB models to describe the distribution

of latent counts. It is possible, however, that some of the excess observations at 20 represent exact consumption by a class of individuals who regularly smoke exactly one pack per day; this is the crux of the Farrell *et al.* [24] model. To check robustness of our conclusions, we extended the underlying ZINB distribution to account for an excess of real consumption of 20; i.e., we added a third term in $f_X(x; \theta)$ to represent the fraction of extra true 20s. Combining this with our heaping (1) model, we estimated the excess exact pack-a-day group to constitute 4.5% of the population. The posterior modes of the non-abstinence proportions are slightly smaller (0.67 and 0.44 for placebo and bupropion, respectively), and the treatment effect in terms of OR of non-abstinence is 0.39. The mean cigarette consumption among the non-abstinent increased by 0.4 in each group (to 12.0 and 12.6), while the ratio of means remained at 1.05. Under the ZIP model, we estimated the pack-a-day fraction to be 0.1%, with the posterior modes of abstinence fraction and smoking rate unchanged.

4. Discussion

We have developed a Bayesian model for heaped data, motivated by cigarette count data from a smoking cessation trial. Our method requires specification of the distribution of the latent outcome variable and the heaping mechanism; estimates of key model parameters may be sensitive to assumptions about either feature. The model is potentially applicable to any substantive area in which data are subject to randomly applied rounding from recall or measurement errors.

Analysis of the clinical trial data using ZI models reveals that bupropion has a substantial effect on the fraction abstaining but not on the smoking rate among the non-abstinent. Thus bupropion may help a smoker to quit, but if he fails to quit it will not help him to otherwise reduce consumption. This finding offers an important insight into bupropion's potential role in combination therapies for smoking cessation.

Inferences from heaped data generally depend on the assumed heaping mechanism. Heitjan and Rubin [18] showed that with age data the placement of the heaping interval, rather than its assumed width, was critical to final inferences. With some heaped variables (e.g., age and blood pressure), there is no question that the heaping is a measurement artifact. With cigarettes it is potentially different, because the fact that cigarettes are sold 20 to a pack suggests that many smokers may indeed consume exactly one pack per day, rendering the heap at 20 a real feature of the distribution on a par with the excess at 0 in a ZI model.

The literature is equivocal. Farrell *et al.* [24] consider the heaps at 5, 10, 15 and 20 to reflect true excess consumption, which they explain as self-rationing. Yet histograms of tobacco by-products in the blood [5], which presumably more accurately reflect real consumption, show no such heaps. Moreover it is unclear why there would be residual heaps at 14, 16, 18, etc., as observed in our data, unless these also are favored counts for self-rationing. Fortunately, some recent studies have acquired both heaped TLFB counts and presumably more accurate counts gleaned from electronic diaries [40]. Analysis of these data should enable us to resolve the question empirically.

If the heaps do represent true consumption, then an efficient strategy for analysis would be to estimate models that account for inflation not only at 0 but at 20 and any other peaks that are believed to represent a likely self-rationing value. Such models are only partially compatible with models that represent the heaping as misreporting, because the inflation of true values must eventually be confounded with the excesses from misreporting. In a sensitivity analysis, we were able to fit a hybrid model that included random misreporting as well as extra exact components at 0 and 20 cigarettes. Estimates were pulled back in the direction of those from unheaped models, but not dramatically so.

Although our model considers cigarette smoking behavior only at one time point, it is common to collect cigarette consumption data throughout treatment and follow-up. Use of the complete data poses serious challenges to modeling. For example, computing the longitudinal model likelihood involves integrating the complete-data density over a set $XG(y)$ that could be large and cumbersome. Models for heaping behavior would also have to be more complex, as some subjects may always report data with the same level of precision, whereas others may report with different levels of precision on different days, depending on the number of cigarettes smoked, the day of the week, and other factors. Further research is necessary to address these questions.

ACKNOWLEDGEMENTS

We gratefully acknowledge the assistance of Caryn Lerman, principal investigator of the bupropion trial, and E. Paul Wileyto, its statistician. The USPHS supported this research under grants 1R01CA116723, 5T32CA093283 and 5P50CA084718.

REFERENCES

1. Collins BN, Wileyto EP, Patterson F, Rukstalis M, Audrain-McGovern J, Kaufmann V, Pinto A, Hawk L, Niaura R, Epstein LH, Lerman C. Gender differences in smoking cessation in a placebo-controlled trial of bupropion with behavioral counseling. *Nicotine & Tobacco Research* 2004;6:27–37. [PubMed: 14982685]
2. Lerman C, Roth D, Kaufmann V, Audrain J, Hawk L, Liu A, Niaura R, Epstein L. Mediating mechanisms for the impact of bupropion in smoking cessation treatment. *Drug & Alcohol Dependence* 2002;67:219–223. [PubMed: 12095672]
3. Wileyto EP, Patterson F, Niaura R, Epstein LH, Brown R, Audrain-McGovern J, Hawk LW, Lerman C. Recurrent event analysis of lapse and recovery in a smoking cessation clinical trial using bupropion. *Nicotine & Tobacco Research* 2005;7:257–268. [PubMed: 16036283]
4. Brown RA, Burgess ES, Sales SD, Whiteley JA, Evans DM, Miller IW. Reliability and validity of a smoking timeline follow-back interview. *Psychology of Addictive Behaviors* 1998;12:101–112.
5. Klesges RC, Debon M, Ray JW. Are self-reports of smoking rate biased? Evidence from the second National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology* 1995;48:1225–1233. [PubMed: 7561984]
6. Lewis-Esquerre JM, Colby SM, Tevyaw TO'L, Eaton CA, Kahler CW, Monti PM. Validation of the timeline follow-back in the assessment of adolescent smoking. *Drug and Alcohol Dependence* 2005;79:33–43. [PubMed: 15943942]
7. Denic S, Khatib F, Saadi H. Quality of age data in patients from developing countries. *Journal of Public Health* 2004;26:168–171. [PubMed: 15284321]
8. Diamond ID, McDonald JW, Shah IH. Proportional Hazards models for current status data: application of the study of differentials in age at weaning in Pakistan. *Demography* 1986;24:607–620. [PubMed: 3803657]
9. Browning M, Crossley TF, Weber G. Asking consumption questions in general purpose surveys. *The Economic Journal* 2003;113:F540–F567.
10. Wolff J, Augustin T. Heaping and its consequences for duration analysis: a simulation study. *Allgemeines Statistisches Archiv* 2003;87:59–86.
11. Locker TE, Mason SM. Digit preference bias in the recording of emergency department times. *European Journal of Emergency Medicine* 2006;13:99–101. [PubMed: 16525239]
12. Pickering RM. Digit preference in estimated gestational age. *Statistics in Medicine* 1992;11:1225–1238. [PubMed: 1509222]
13. Savitz DA, Terry JW Jr, Dole N, Thorp JM Jr, Siega-Riz AM, Herring AH. Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination. *American Journal of Obstetrics and Gynecology* 2002;187:1660–1666. [PubMed: 12501080]

14. de Lusignan S, Belsey J, Hague N, Dzregah B. End-digit preference in blood pressure recordings of patients with ischaemic heart disease in primary care. *Journal of Human Hypertension* 2004;18:261–265. [PubMed: 15037875]
15. Thavarajah S, White WB, Mansoor GA. Terminal digit bias in a specialty hypertension family practice. *Journal of Human Hypertension* 2003;17:819–822. [PubMed: 14704725]
16. Roberts JM, Brewer DD. Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics* 2001;28:887–896.
17. Crawford SL, Johannes CB, Stellato RK. Assessment of digit preference in self-reported year at menopause: choice of an appropriate reference distribution. *American Journal of Epidemiology* 2002;156:676–683. [PubMed: 12244037]
18. Heitjan DF, Rubin DB. Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association* 1990;85:304–314.
19. Ridout MS, Morgan BJT. Modeling digit preference in fecundability studies. *Biometrics* 1991;47:1423–1433. [PubMed: 1786326]
20. Dellaportas, P.; Stephens, DA.; Smith, AFM.; Guttman, I. A comparative study of perinatal mortality using a two component mixture model. In: Berry, DA.; Stangl, DK., editors. *Bayesian Biostatistics*. New York: Dekker; 1996. p. 601–616.
21. Wright DE, Bray I. A mixture model for rounded data. *The Statistician* 2003;52:3–13.
22. Ramsay JO. Monotone regression splines in action (with discussion). *Statistical Science* 1988;3:425–461.
23. Shryock, HS.; Siegel, JS., et al. *The Methods and Materials of Demography*. Vol. Volumes 1 and 2. Washington DC: U. S. Department of Commerce, Bureau of the Census; 1973.
24. Farrell, L.; Fry, TRL.; Harris, MN. “A pack a day for twenty years”: Smoking and cigarette pack sizes. Research paper number 887, Department of Economics, University of Melbourne. 2003.
25. Heitjan DF, Rubin DB. Ignorability and coarse data. *Annals of Statistics* 1991;19:2244–2253.
26. Heitjan DF. Ignorability and coarse data: some biomedical examples. *Biometrika* 1993;49:1099–1109.
27. Heitjan DF. Ignorability in general incomplete-data models. *Biometrika* 1994;81:701–708.
28. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992;34:1–14.
29. Ridout M, Hinde J, Demétrio CGB. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* 2001;57:219–223. [PubMed: 11252601]
30. Hurt RD, Sachs DPL, Glover ED, Offord KP, Johnston JA, Dale LC, et al. A comparison of sustained-release bupropion and placebo for smoking-cessation. *New England Journal of Medicine* 1997;337:1195–1202. [PubMed: 9337378]
31. Jorenby DE, Leischow SJ, Nides MA, Rennard SI, Johnston JA, Hughes AR, et al. A controlled trial of sustained-release bupropion, a nicotine patch, or both for smoking cessation. *New England Journal of Medicine* 1999;340:685–691. [PubMed: 10053177]
32. Piasecki TM. Relapse to smoking. *Clinical Psychology Review* 2006;26:196–215. [PubMed: 16352382]
33. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis*. Vol. second edition. Boca Raton, FL: Chapman & Hall/CRC; 2003.
34. Tanner, MA. *Tools for Statistical Inference*. Vol. second edition. New York: Springer; 1993.
35. Press, WH.; Flannery, BP.; Teukolsky, SA.; Vetterling, WT. *Numerical recipes: The Art of Scientific Computing*. Cambridge: Cambridge University Press; 1988.
36. Dennis, JE., Jr; Schnabel, RB. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall; 1983.
37. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995;90:773–795.
38. Gelman A, Meng XL. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* 1998;13:163–185.

39. Kass RE, Greenhouse JB. A Bayesian Perspective. Comment on: investigating therapies of potentially great benefit. *Statistical Science* 1989;4:310–317.
40. Shiffman S, Hickcox M, Paty JA, Gnys M, Kassel JD, Richards TJ. Progression from a smoking lapse to relapse: Prediction from abstinence violation effects, nicotine dependence, and lapse characteristics. *Journal of Consulting and Clinical Psychology* 1996;64:993–1002. [PubMed: 8916628]

Figure 1.
Reported cigarette counts on day 56 (EOT).

Figure 2.
Reported cigarette count Y as a function of the underlying count X and the rounding behavior G in our model of reported cigarette counts.

Figure 3.
Square-root scale cigarette count residuals.

Figure 4. 95% probability intervals for the fraction of non-abstainers and mean consumption among non-abstainers. × indicates the posterior mode.

Figure 5.
Estimated rounding behavior G under the ZINB heaped (1) model.

Table 1

Bias in the mean from heaping Poisson data.

True Mean	Probability of Heaping			
	0.1	0.3	0.5	0.7
10	-0.17	-0.50	-0.83	-1.16
15	0.26	0.79	1.32	1.85
20	0.02	0.07	0.11	0.15
25	-0.14	-0.41	-0.68	-0.96
30	0.05	0.15	0.25	0.34

Table II
Twice the natural logarithm of the Bayes factors in the bupropion data (column versus row).

	Poisson			Negative binomial		
	Simple	ZI	ZI, heaped(1)	Simple	ZI	ZI, heaped(1)
Poisson						
Simple		7997	9123	9628	9991	10705
ZI			1125	1631	1993	2708
ZI, heaped (1)				506	868	1583
ZI, heaped (2)			-13.3			
NB						
Simple					362	1077
ZI						715
ZI, heaped (2)						5.7

Table III

Mean cigarette consumption (given non-abstinence for ZI models): Posterior mode [95% probability interval].

Arm	Model/Distribution	Poisson	Negative binomial
Placebo	Simple	7.9 [7.7,8.2]	7.9 [6.7,9.5]
	ZI	13.0 [12.6,13.4]	12.7 [11.6,13.9]
	ZI, heaped (1)	11.2 [10.8,11.7]	11.6 [10.3,13.1]
	ZI, heaped (2)	11.4 [11.0,11.9]	12.6 [11.4,13.8]
Bupropion	Simple	5.5 [5.3,5.7]	5.5 [4.6,6.5]
	ZI	13.6 [13.1,14.1]	13.3 [12.0,14.8]
	ZI, heaped (1)	11.8 [11.2,12.4]	12.2 [10.6,13.8]
	ZI, heaped (2)	12.0 [11.4,12.6]	13.2 [11.8,14.7]
Ratio of means	Simple	0.69 [0.66,0.72]	0.69 [0.54,0.87]
	ZI	1.04 [0.99,1.09]	1.05 [0.92,1.20]
	ZI, heaped (1)	1.05 [0.99,1.11]	1.05 [0.90,1.22]
	ZI, heaped (2)	1.05 [0.98,1.11]	1.05 [0.91,1.20]

Table IV

Fraction non-abstinent: Posterior mode [95% probability interval].

Arm	Model/Distribution	Poisson	Negative binomial
Placebo	ZI	0.61 [0.57,0.65]	0.62 [0.58,0.67]
	ZI, heaped (1)	0.67 [0.62,0.72]	0.70 [0.63,0.77]
	ZI, heaped (2)	0.61 [0.57,0.65]	0.62 [0.58, 0.67]
Bupropion	ZI	0.40 [0.36,0.45]	0.41 [0.37,0.46]
	ZI, heaped (1)	0.44 [0.39,0.48]	0.46 [0.40,0.52]
	ZI, heaped (2)	0.40 [0.36,0.45]	0.41 [0.37, 0.46]
Odds ratio	ZI	0.43 [0.33,0.55]	0.42 [0.33,0.54]
	ZI, heaped (1)	0.38 [0.28,0.51]	0.36 [0.26,0.50]
	ZI, heaped (2)	0.43 [0.33,0.55]	0.42 [0.32,0.55]

Table V
Sensitivity to the assumed prior: Posterior modes of selected parameters in ZINB heaped (1)

Prior	Posterior mode					
	θ_1		θ_2			
	Placebo	Bupropion	Placebo	Bupropion	OR	Ratio of means
$\beta_2 \sim N(0, 10^2)$	0.70	0.46	11.6	12.2	0.36	1.05
$\beta_1 \sim N(0, 10^2)$	0.70	0.46	11.6	12.2	0.36	1.05
$\eta_1 \sim N(2.996, 10^2)$	0.70	0.46	11.6	12.2	0.36	1.05
$\beta_1 \sim t_1(0, 0.459^2)$, and $\beta_2 \sim t1(0, 1.099^2)$	0.68	0.46	11.8	12.3	0.39	1.05
$\eta_1 \sim t_1(2.3, 0; 0.459^2)$, and $\eta_2 \sim t1(0, 1.099^2)$	0.71	0.46	11.6	12.1	0.36	1.05
Normal with $\sigma_N = 2$	0.72	0.47	11.3	11.9	0.35	1.05
Normal with $\sigma_N = 0.6$	0.81	0.55	10.3	10.6	0.28	1.04
$\beta_1, \beta_2 \sim U[-50, 50]$	0.70	0.46	11.6	12.2	0.36	1.05
$\eta_1, \eta_2 \sim U[-10, 10]$	0.70	0.46	11.6	12.2	0.36	1.05
$\gamma_0, \gamma_1, \gamma_2, \gamma_3 \sim U[-20, 20]$	0.70	0.46	11.6	12.2	0.36	1.05
$\theta_3 \sim \varepsilon(1)$	0.71	0.47	11.4	12.0	0.35	1.05