Model-based imputation of latent cigarette counts using data from a calibration study

SANDRA D. GRIFFITH,¹ SAUL SHIFFMAN,² YIMEI LI³ & DANIEL F. HEITJAN^{4,5}

¹ Flatiron Health

² Psychology, University of Pittsburgh

³ Biostatistics & Epidemiology, University of Pennsylvania

⁴ Statistical Science, Southern Methodist University

⁵ Clinical Sciences, University of Texas Southwestern

Key words

addiction, nicotine, tobacco, statistics, methodology

Correspondence

Sandra Griffith, Flatiron Health,
New York, NY 10012, USA.
Email: sgriffith@flatiron.com

Received 8 January 2014;
revised 15 May 2014;
accepted 27 June 2014

Abstract

In addition to dichotomous measures of abstinence, smoking studies may use daily cigarette consumption as an outcome variable. These counts hold the promise of more efficient and detailed analyses than dichotomous measures, but present serious quality issues – measurement error and heaping – if obtained by retrospective recall. A doubly-coded dataset with a retrospective recall measurement (timeline followback, TLFB) and a more precise instantaneous measurement (ecological momentary assessment, EMA) serves as a calibration dataset, allowing us to predict EMA given TLFB and baseline factors. We apply this model to multiply impute precise cigarette counts for a randomized, placebo-controlled trial of bupropion with only TLFB measurements available. To account for repeated measurements on a subject, we induce correlation in the imputed counts. Finally, we analyze the imputed data in a longitudinal model that accommodates random subject effects and zero inflation. Both raw and imputed data show a significant drug effect for reducing the odds of non-abstinence and the number of cigarettes smoked among non-abstainers, but the imputed data provide efficiency gains. This method permits the analysis of daily cigarette consumption data previously deemed suspect due to reporting error and is applicable to other self-reported count data sets for which calibration samples are available. Copyright © 2015 John Wiley & Sons, Ltd.

Introduction

Although smoking cessation clinical trials typically use as primary outcome the dichotomous assessment of abstinence versus smoking, there is also interest in assessing how much people smoke. For example, many trials of smoking

reduction require the assessment of cigarette consumption in much finer detail (Hughes and Carpenter, 2005) – specifically, the actual numbers of cigarettes smoked.

It is common to record daily cigarette counts using timeline followback (TLFB), a form of structured retrospective recall (Sobell and Sobell, 1992). The resultant

counts, however, typically contain excess observations at multiples of 20, 10, and 5. *Heaping*, a type of measurement error characterized by numbers reported with varying levels of precision, is ubiquitous in self-reported count data (Heitjan and Rubin, 1990; Jukic *et al.*, 2008; Komlos, 1999; Wiederman, 1997; Wutich, 2009), including those collected in psychological and substance abuse research studies (Huttenlocher *et al.*, 1990; Matt and Jo, 1994). While heaping provides one explanation for this reporting pattern, it is also possible that features in the distributions of the measured quantities are responsible for the heaping. For example, Farrell *et al.* (2003) proposed that smokers ration daily cigarette consumption into full- or half-pack intervals, suggesting that heaps at those values are a genuine feature of smoking behavior rather than a bias in self-report; however, evidence from nicotine by-product measures, which lack any corresponding heaps, suggest that heaping error is the more likely explanation (Klesges *et al.*, 1995).

While it is in principle preferable to use all the information available, analyses based on count data that are heavily heaped can be unreliable. Wang and Heitjan (2008) demonstrated that simple random heaping can bias estimates of mean daily cigarette consumption by as much as 20%, with the exact bias depending on both the true mean and the extent of heaping. Heaping also leads to efficiency loss (Heitjan, 1989; Schneeweiss *et al.*, 2010).

Ecological momentary assessment (EMA) studies record cigarettes smoked in real time. While such data are often analyzed for the details of smoking episodes, they can also be used to generate more precise cigarette counts. Although not error-free, EMA data avoid the retrospective reporting errors and corresponding heaping seen in TLFB; they also exhibit higher correlation with nicotine metabolites (Shiffman, 2009b). As few smoking studies use EMA, it would be useful to leverage the relatively more precise data collected by EMA in one study to data from other studies by modeling the relationship between potentially-heaped TLFB data and true, unheaped cigarette counts, for which we use EMA as a proxy.

To effect such an analysis, we employ a dataset, collected by Shiffman (2009b), that contains both retrospectively-reported (TLFB) and real-time (EMA) daily counts. This doubly-coded dataset serves as a calibration tool for modeling the relationship between heaped TLFB and true counts, with EMA serving as the latter's surrogate. We apply the results of our predictive model to impute latent counts in a randomized, placebo-controlled trial of bupropion (Li *et al.*, 2011; Wileyto *et al.*, 2005) in which daily cigarette counts were measured only by TLFB.

Others have addressed heaping in cigarette smoking applications, typically combining models for the heaping

behavior with parametric models for the underlying counts; these analyses assess the extent of heaping and its impact on bias and efficiency (Bar and Lillard, 2012; Wang and Heitjan, 2008; Wang *et al.*, 2012). To our knowledge this is the first analysis to attempt the imputation of precise values for the heaped daily cigarette counts.

Methods

Participants

Calibration data

Participants were 236 adult smokers enrolled in a smoking cessation study (Shiffman, 2009b). For 14 days prior to a designated quit date, investigators collected two measurements of daily cigarette consumption: one by EMA and the other TLFB. The EMA instrument required participants to log each cigarette on an electronic diary (ED) just before smoking it. Each evening, participants had an opportunity to record any cigarettes they failed to log during the day. The ED prompted participants at four to five randomly selected times per day; participants responded to 91% of such prompts within the two minutes allowed, indicating that they were carrying and attending to the device.

Imputation data

We analyze a dataset describing daily cigarette consumption in the treatment phase of a randomized, placebo-controlled smoking cessation trial of bupropion, containing daily cigarette counts only recorded by TLFB (Li *et al.*, 2011; Wileyto *et al.*, 2005). The trial recorded up to 57 TLFB measurements per subject on 1108 subjects. We observe heaping at multiples of 5 and 10 (Figure 1, bottom). Unlike the calibration dataset, the imputation dataset includes treatment-phase measurements, and hence a marked excess of zeroes (59%) from days of successful abstinence.

Predictive model

We use the calibration data to fit a log-linear mixed-effects model predicting EMA counts, assuming a Poisson distribution conditional on TLFB counts and other factors available in both datasets. Predictors include the observed TLFB count, log-transformed for interpretability, as we expect the reported count to be associated with the true count; three indicator variables for the TLFB-reported count's divisibility by 20, 10, and 5, respectively, based on our observation that participants exhibit the greatest preference for reporting these multiples, as compared to other types of numbers (i.e. multiples of two not also divisible by 10 and odd numbers not divisible by five); the

Figure 1. Histograms of daily cigarette consumption in the calibration dataset, as measured by timeline followback (TLFB) and ecological momentary assessment (EMA), and in the imputation dataset, as measured by TLFB. Imputation dataset measurements displayed here include only non-zero counts to allow comparison of the three data sources on the same scale; zero counts constitute 59% of measurements in the imputation dataset.

within-subject standard deviation (SD) of TLFB-reported counts to account for the association between heaping and reporting invariance; and person-level variables that could affect the EMA–TLFB relationship: Fagerström Test for Nicotine Dependence (FTND) score, sex, and race. Independent normally distributed random intercepts at the subject and observation levels represent additional heterogeneity between subjects and overdispersion, respectively. For computational convenience, we chose a normal distribution for the overdispersion term; however, results are generally similar with a negative binomial model, which is equivalent to assuming a gamma distribution for the random effects (Breslow, 1990).

We employ Bayesian methods to estimate our predictive model, because they provide convenient access to samples from the posterior distribution of the parameters drawn from during the imputation procedure. We assign vague proper priors to the fixed effect parameters; for the variance components, we assign inverse-Wishart priors with low degree-of-belief parameters and centrality parameters based on empirical estimates of the variability in EMA. We employ a Markov Chain Monte Carlo (MCMC) approach to estimate all parameters, with 1000 iterations

remaining for each chain after a burn-in of 3000 iterations and a thinning interval of 1:10 to reduce autocorrelation. Running two chains with different starting values allows us to check the mixing properties of the two sequences and ensure approximate convergence through graphical diagnostics and Gelman–Rubin statistics (Gelman and Rubin, 1992). We implement the MCMC estimation in R version 2.15.1 (R Development Core Team, 2012) using the MCMCglmm package (Hadfield and Kruuk, 2010).

Imputation

Based on experimental evidence for a cognitive distinction between zero and other even numbers (Nuerk *et al.*, 2004), we assume that zero counts in the imputation data are precise and represent an abstinent state. For the non-zero counts, we created 10 complete datasets by the following procedure:

- (1) Draw a vector of parameters from the approximate posterior distribution (obtained from the MCMC estimation for the predictive model).
- (2) For each subject in the imputation dataset, draw a random intercept from a normal distribution with mean zero and variance set to its value drawn in Step 1.

- (3) For each observation within each subject, draw an overdispersion error from a normal distribution with mean zero and variance set to its value drawn in Step 1.
- (4) Compute a linear predictor using the covariates, parameter estimate draws from Step 1, and random effects from Steps 2 and 3.
- (5) Randomly sample precise, EMA-like counts from a Poisson distribution with mean equal to the exponentiated linear predictor from Step 4.

Repeating the procedure 10 times gives 10 imputation datasets amenable to analysis by complete-data methods.

Complete-data analysis

We fit a regression model to each of the 10 imputed datasets. Because the imputation data are from the quit period, they contain excess zeroes which must be accounted for in the regression model. We favor a model that directly describes the probability of abstinence – a clinically relevant quantity. Within the class of zero-inflated models, the *hurdle model* employs a two-stage estimation process and provides greater computational reliability over zero-inflated Poisson or negative binomial models (Min and Agresti, 2005). The hurdle model estimates the probability of smoking versus abstinence by logistic regression and the distribution of counts given non-abstinence by a zero-truncated Poisson. Both model components include observation-level predictors representing clinically-important features used in previous analyses (treatment group, sex, race, FTND score, and week). To accommodate repeated measures, we include correlated random effects in both models, allowing a relationship between the probability of smoking and mean cigarette count on smoking days. We fit the model using Proc NLMIXED in SAS version 9.3 (SAS Institute, Cary, NC) and summarize results over 10 imputations (Rubin, 1987) using Proc MIANALYZE. For comparison, we fit the same model using the raw reported cigarette counts. In cases where subject-level variability is high, conditional estimates from mixed-effects models can differ substantially from marginal, population-averaged estimates. Thus, we re-scale conditional estimates to average over the distribution of the random effects and provide approximate marginal estimates for comparison (Diggle *et al.*, 2004).

Imputation diagnostics

To identify residual heaping in the imputed data, we perform graphical posterior predictive checks (Gelman *et al.*, 2005). We evaluate histograms of imputed values for each of several imputations in terms of an implicit smooth reference distribution. Although draws from

smooth Poisson distributions form the basis of the imputations, the mixture of Poisson distributions allows for a potential preference for rounded values. If any ambiguity remains after these graphical checks, one can also perform quantitative posterior predictive checks on heaping-related metrics, e.g. the proportions of observations divisible by 5, 10, or 20. Expected proportions of those quantities in the absence of any digit preference form the implicit reference for this comparison. Calculating the intra-class correlation coefficient (ICC) for each imputed dataset provides a check that the imputations retain appropriate within-subject correlation.

To further assess the performance of the imputation method, we applied it to the calibration dataset where a proxy for the precise measurement is known and serves as an explicit reference distribution. Specifically, we used the predictive model fit with the calibration dataset to impute EMA-like counts from observed calibration TLFB counts, as if the true EMA counts were not known. To diagnose any lack of fit in the predictive or imputation models, we performed several graphical posterior predictive checks. First, we compared histograms of imputed datasets to the reference dataset to detect discrepancies in location and variability of the distributions. Next, we directly compared the imputed data points to the reference EMA counts in a scatterplot to detect any departures from the diagonal identity line.

Results

Cigarette consumption

Our analysis of the calibration data excluded 27 person-days on which EDs malfunctioned, 11 person-days with TLFB or EMA counts of zero, and 118 person-days for which cigarette counts were unavailable. This left 3148 pairs on 236 subjects (mean = 13.3 days/subject out of a possible 14 days). The sample was majority female (59%) and White (92%) with mean age 43.6 (SD = 9.9) years and a median FTND score of six (interquartile range: 5–8; scale ranges from zero to 10 with higher numbers indicating greater dependence). The top two panels of Figure 1 show histograms of their cigarette counts. The TLFB version exhibits marked heaps at multiples of 5, 10, and 20 cigarettes per day, as compared to the smooth distribution of EMA. Although EMA counts likely contain some measurement error, the smooth distribution seen here, together with their correlation with nicotine metabolites (Shiffman, 2009b), suggest an absence of heaping error.

The calibration data exhibit a high correlation between repeated measurements within a subject, quantified by an

Table 1. Summary of the posterior distribution of the parameters for the Poisson log-linear mixed-effects model, fit using the calibration data to predict EMA-measured cigarette counts

Parameter	Mean	SD	2.5th centile	97.5th centile	pMCMC
Intercept	1.035	0.186	0.686	1.408	<0.001
log(TLFB)	0.477	0.032	0.413	0.538	<0.001
TLFB divisible by 20	0.000	0.024	−0.046	0.045	0.992
TLFB divisible by 10	0.010	0.023	−0.036	0.057	0.678
TLFB divisible by 5	−0.008	0.019	−0.046	0.028	0.650
FTND	0.040	0.017	0.006	0.075	0.022
Sex	0.013	0.075	−0.135	0.165	0.862
Race	0.202	0.145	−0.078	0.476	0.160
SD (TLFB)	0.011	0.016	−0.019	0.042	0.494
σ_u	0.292	0.027	0.243	0.348	—
σ_e	0.062	0.002	0.058	0.066	—

Note: σ_u and σ_e are the square roots of the subject-specific and observation-level random effect variance components, respectively. We calculated the pMCMC statistic as twice the minimum of the proportion of posterior estimates either above or below zero.

Table 2. Complete-data analysis estimates from the mixed-effects hurdle model

Parameter	Raw		Imputed	
<i>Binary model</i>	<i>OR</i>	<i>Adj OR</i>	<i>OR</i>	<i>Adj OR</i>
(Smoking)	(95% CI)	(95% CI)	(95% CI)	(95% CI)
Drug	0.05 (0.02, 0.10)	0.41 (0.33, 0.51)	0.05 (0.02, 0.13)	0.49 (0.39, 0.60)
Week	1.51 (1.49, 1.53)	1.13 (1.12, 1.13)	1.53 (1.51, 1.56)	1.11 (1.11, 1.12)
Sex (female)	6.25 (3.02, 12.95)	1.70 (1.38, 2.11)	9.53 (3.89, 23.32)	1.74 (1.4, 2.17)
Race (white)	0.64 (0.23, 1.73)	0.88 (0.66, 1.17)	0.83 (0.25, 2.81)	0.96 (0.71, 1.29)
FTND	1.29 (1.09, 1.52)	1.08 (1.03, 1.13)	1.03 (0.84, 1.26)	1.01 (0.96, 1.06)
<i>Count model</i>	<i>RR</i>		<i>RR</i>	
	(95% CI)		(95% CI)	
Drug	0.58 (0.48, 0.69)	—	0.82 (0.76, 0.89)	—
Week	1.05 (1.05, 1.05)	—	1.03 (1.03, 1.04)	—
Sex (female)	1.25 (1.04, 1.50)	—	1.12 (1.03, 1.22)	—
Race (white)	1.15 (0.90, 1.47)	—	1.08 (0.96, 1.21)	—
FTND	1.11 (1.07, 1.16)	—	1.08 (1.04, 1.12)	—
<i>Variance components</i>	<i>Estimate (SE)</i>		<i>Estimate (SE)</i>	
σ_{u1}	5.59 (0.18)	—	6.69 (0.31)	—
σ_{u2}	1.30 (0.48)	—	0.56 (0.04)	—
ρ	0.77 (0.02)	—	0.81 (0.02)	—

Note: Estimates for the binary and count models are expressed as odds ratio (ORs) and rate ratios (RRs), respectively. The reported timeline followback (TLFB) counts in the imputation dataset form the basis for the raw estimates. Imputed estimates combine information from 10 sets of imputed ecological momentary assessment (EMA) counts. Adjusted (Adj) ORs average over the distribution of the random effects and give an approximate marginal (population-averaged) interpretation of the estimates for the binary model; no approximation is needed for the count model, where RRs have both a conditional and marginal interpretation. σ_{u1} and σ_{u2} are the square roots of the subject-specific and observation-level random effect variance components, respectively; we denote the correlation between these two random effects as ρ .

ICC of 0.83 (95% confidence interval [CI]: 0.80, 0.86) for TLFB measurements and 0.76 (95% CI: 0.72, 0.80) for EMA measurements. Many subjects report the same TLFB observation (most commonly 20, presumably a consequence of heaping) day after day, whereas EMA values typically vary within subjects; this may explain the higher ICC for TLFB measurements.

The imputation data include 1108 subjects and 61,187 measurements, 41% of them non-zero; for an additional 1969 person-days, cigarette counts were unavailable. Removing 951 observations with missing covariates or outcomes left 60,236 observations on 1066 subjects (mean = 52.2 days/subject out of a possible 57 days). The sample was majority female (52%) and White (85%) with a median FTND score of seven (interquartile range: 5–8). The non-zero TLFB counts in the imputation dataset exhibit within-subject correlation similar to the calibration dataset, sharing the same ICC point estimate (0.83) with a slightly narrower 95% CI (0.82, 0.84). The within-subject correlation of TLFB measurements from the full imputation dataset – both zero and non-zero counts – is slightly higher (ICC = 0.90, 95% CI: 0.88, 0.92).

Predictive model

Table 1 presents posterior means, 95% credible intervals, and MCMC *p*-values for the predictive model fit to the calibration dataset. TLFB count strongly predicts EMA count, with each increase in log TLFB producing an increase of 0.48 in expected log EMA after covariate adjustment. FTND was a significant, if weak, predictor of log EMA count. Graphical convergence diagnostics suggest adequate mixing, confirmed by upper bounds close to unity for all Gelman–Rubin statistics.

Complete-data analysis

Table 2 displays the complete-data analysis of the imputation dataset. We compare raw TLFB-based estimates, subject to heaping, to the combined estimates from 10 imputed datasets. Consistent with previous analyses (Li *et al.*, 2011; Wileyto *et al.*, 2005), subjects in the bupropion group have a statistically significant reduction in the odds of smoking (odds ratio [OR]: 0.05, Raw 95% CI: 0.02, 0.10), controlling for covariates. The estimated drug effect retains similar magnitude and significance using

Table 3. Multiple imputation summary statistics

	Variance						
Parameter	Between	Within	Total	DF	RI Var	FMI	RE
<i>Binary model (Smoking)</i>							
Intercept	0.002	0.789	0.792	1058	0.003	0.003	1.000
Drug	0.001	0.198	0.199	1053	0.005	0.005	0.999
Week	0.000	0.000	0.000	1008	0.018	0.018	0.998
Sex (female)	0.006	0.203	0.209	938	0.030	0.030	0.997
Race (White)	0.002	0.382	0.384	1055	0.005	0.004	1.000
FTND	0.001	0.010	0.011	678	0.069	0.066	0.993
σ_{u1}	0.028	0.064	0.094	78	0.473	0.336	0.967
<i>Count model</i>							
Intercept	0.027	0.007	0.036	12	4.539	0.841	0.922
Drug	0.000	0.002	0.002	790	0.052	0.050	0.995
Week	0.000	0.000	0.000	15	3.077	0.781	0.928
Sex (female)	0.000	0.002	0.002	499	0.103	0.095	0.991
Race (white)	0.000	0.003	0.003	480	0.107	0.099	0.990
FTND	0.000	0.000	0.000	18	2.285	0.724	0.933
σ_{u2}	0.001	0.001	0.002	18	2.110	0.707	0.934
ρ	0.000	0.000	0.001	106	0.372	0.283	0.973

Note: σ_{u1} and σ_{u2} are the square roots of the subject-specific and observation-level random effect variance components, respectively; we denote the correlation between these two random effects as ρ . Between-imputation, within-imputation, and total variability, in conjunction with imputation degrees of freedom (DF), dictate the remaining summary results. Relative increase in variance (RI Var) estimates the impact of missing data on the total variance. The fraction of missing information (FMI) estimates the proportion of information lost for each parameter estimate due to missing data. The relative efficiency (RE) quantifies the loss in efficiency due to imputing 10 complete datasets, compared to an infinite number of imputations.

combined information from the 10 imputed datasets. The imputation procedure alters only cigarette counts on non-abstinent days; thus, we would not expect large changes in the binary component of the model. The count model attenuates the estimated drug effect on cigarette counts among non-abstainers for the imputed data (rate ratio [RR] [95% CI]: 0.82 [0.76, 0.89]), as compared to the raw data (RR [95% CI]: 0.58 [0.48, 0.69]). The imputed data estimate reflects an overall increase in precision, with reductions of up to 59% in the width of CIs for the RRs. High estimates for the variance components in both the raw (binary model: 5.59 [standard error, SE: 0.18]; count model: 1.30 [SE: 0.48]) and imputed data (binary model: 6.69 [SE: 0.31]; count model: 0.56 [SE: 0.04]) indicate substantial heterogeneity between subjects, particularly in predisposition to smoke. Because of this heterogeneity, the approximate marginal ORs in the binary model are attenuated compared to their conditional counterparts.

Table 3 summarizes the multiple imputation results obtained by Proc MIANALYZE. Low between-imputation variability for all parameter estimates indicates that the imputation procedure produced consistent results upon repeated applications. The high relative efficiency (RE) for all parameter estimates suggests that using 10 imputations provides adequate efficiency. For most parameters in the binary model, the estimated fraction of missing information (FMI) is low, largely dictated by a favorable ratio of between- to within-imputation variability. The variance component for the binary model's subject-specific random intercept, σ_{ui} , presents a notable exception with an estimated 33.6% of information missing. Although some parameter estimates in the count model exhibit a high FMI, this is likely an artifact of the extremely low magnitude of between- and within-imputation variability. The degree-of-freedom adjustment produces only a minor change in *p*-values and no change in overall conclusions.

Figure 2. Ten imputed datasets, depicted in grey, superimposed on non-zero TLFB values from the imputation dataset in black. All histogram bins have width equal to one. The smooth appearance of the imputed datasets suggests that the imputation eliminates heaping.

Imputation diagnostics

Figure 2 displays the distributions of the 10 imputed datasets, with the original heaped TLFB values as comparison. The smooth appearance of the imputed datasets provides graphical evidence for the absence of heaping, precluding the need for quantitative posterior predictive checks for heaping in the imputed data. The ICC over 10 imputed datasets averages 0.73 and ranges from 0.70 to 0.76, consistent with what one would expect given the relationship of within-subject correlation between TLFB and EMA in the calibration dataset and the estimated ICC in the imputation dataset.

Although it is clear that the imputed data do not exhibit heaping, Figure 2 provides incomplete information about the accuracy of the imputations. Figure 3 examines the marginal distribution of the imputed EMA-like values, with the observed EMA included as an explicit reference. The distributions reveal between-imputation variability in both the mean and variance,

with several of the imputations exhibiting comparable mean and variance to the true values. Figure 4 examines the joint distribution of the imputed and observed EMA, demonstrating good agreement overall, despite some attenuation for large values of imputed EMA.

Discussion

We have demonstrated a method for imputing daily cigarette counts using data collected in a calibration study. The method accounts for measurement error due to heaping in studies where counts are available only by TLFB measurements. Estimates of key parameters in the count model differ substantially under the imputation and raw-data analyses. In particular, the imputation shrinks every RR parameter toward one – most notably the drug effect, whose 95% CI under the imputation model does not overlap with that from the raw-data analysis. Although these changes do not affect statistical significance, the large difference in effect sizes suggests that the heaping error can

Figure 3. Histograms of 10 imputed EMA datasets, depicted in grey, superimposed on observed EMA values from the calibration dataset in black. All histogram bins have width equal to one.

Figure 4. Distribution of imputed EMA values from a single imputed dataset as a function of the corresponding observed EMA values. The dashed diagonal line of equality represents agreement between the two values. Jittering and transparency (factor of 15) were applied to alleviate over-plotting.

induce substantial bias. Estimated variance components from the imputation analysis also differ from the raw-data analyses. Improper estimation of variance components – reflecting the tendency of some subjects to report the same, heaped count regardless of the number of cigarettes actually smoked – can create bias and inefficiency in mixed models analysis. Finally, we observe smaller standard errors for most count model parameters under imputation, consistent with the general observation that estimates based on rounded data are less efficient than those based on unrounded data (Heitjan, 1989; Schneeweiss *et al.*, 2010); this could be pivotal in examples with smaller sample sizes or more modest drug effects.

Our method relies on several assumptions regarding the TLFB and EMA data. Although we treat EMA as a proxy for the latent true values, these measurements are not entirely error-free, which poses a potential limitation to our findings. The high compliance rate to EMA prompts and strong correlation with smoking bio-markers (Shiffman, 2009a), however, suggest that any errors in EMA data are modest compared to those in TLFB data.

The smooth distribution of EMA measurements indicate that any errors present are distinct from heaping error. Although TLFB counts were on average 3.2 cigarettes higher than EMA, on one-third of days EMA counts exceeded TLFB reports, suggesting that discrepancies between the two measures are not merely a failure to record cigarettes using EMA (Griffith *et al.*, 2009). It is possible that the participants' smoking behavior was affected by the EMA monitoring, but previous analyses of these data showed only small reductions in smoking frequency over the course of the monitoring period (0.3 cigarettes per day) and no significant changes in carbon monoxide levels over the same period (Shiffman *et al.*, 2002).

Lastly, our method assumes that the distribution of heaping errors is the same in both datasets after adjustment for observed confounders; because we possess both EMA and TLFB data only in the calibration set, we cannot test this hypothesis directly. Diagnostics on the calibration data suggest that, at a minimum, the difference in cigarette consumption levels between the *ad lib* and treatment-phase smoking periods does not contribute to lack of fit. It is also possible that unmeasured confounders for which

we could not adjust may affect the heaping mechanism. We acknowledge the discrepancy between the calibration and imputation datasets as a potential limitation to our results.

We observe that the imputation of precise counts informed by models relating heaped to precise counts can reduce both bias and inefficiency in analysis. This provides further evidence that numeric counts obtained by retrospective recall, while potentially adequate for gross

comparisons among subjects, may not be suitable for settings requiring precise assessments of cigarette consumption or assessment of changes in consumption over time. In these cases, targeted deployment of EMA methods could unlock information in existing count data, heretofore deemed unusable due to heaping. These methods are moreover potentially applicable to other areas of psychiatric and substance use research that make use of self-reported count data.

References

- Bar H.Y., Lillard D.R. (2012) Accounting for heaping in retrospectively reported event data – a mixture-model approach. *Statistics in Medicine*, **31**(27), 3347–3365. DOI: 10.1002/sim.5419.
- Breslow N. (1990) Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association*, **85**(410), 565–571.
- Diggle P., Heagerty P., Liang K., Zeger S. (2004) *Analysis of Longitudinal Data*, 2nd edition, Oxford: Oxford University Press.
- Farrell L., Fry T., Harris M. (2003) "A pack a day for twenty years": Smoking and Cigarette Pack Sizes, Research Paper 887, Melbourne, Department of Economics, University of Melbourne. <http://ideas.repec.org/p/mlb/wpaper/887.html>.
- Gelman A., Rubin D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**(4), 457–472. DOI: 10.1214/ss/1177011136.
- Gelman A., Van Mechelen I., Verbeke G., Heitjan D.F., Meulders M. (2005) Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics*, **61**, 74–85. DOI: 10.1111/j.0006-341X.2005.031010.x.
- Griffith, S.D., Shiffman, S., Heitjan, D.F. (2009) A method comparison study of timeline followback and ecological momentary assessment of daily cigarette consumption. *Nicotine & Tobacco Research*, **11**(11), 1368–1373. DOI: 10.1093/ntn/ntp150.
- Hadfield J.D., Kruuk L. (2010) MCMC methods for multi-response generalised linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, **33**(2), 1–22. DOI: 10.1111/j.1533-7168.2010.00312.x.
- Heitjan D.F. (1989) Inference from grouped continuous data: a review. *Statistical Science*, **4**(2), 164–179. DOI: 10.1214/ss/1177012601.
- Heitjan D.F., Rubin D.B. (1990) Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, **85**(410), 304–314. DOI: 10.2307/2289765.
- Hughes J.R., Carpenter M.J. (2005) The feasibility of smoking reduction: an update. *Addiction*, **100**(8), 1074–1089.
- Huttenlocher J., Hedges L., Bradburn N. (1990) Reports of elapsed time: bounding and rounding processes in estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**(2), 196–213. DOI: 10.1037/0278-7393.16.2.196.
- Jukic A.M.Z., Weinberg C.R., Wilcox A.J., McConaughy D.R., Hornsby P., Baird D.D. (2008) Accuracy of reporting of menstrual cycle length. *American Journal of Epidemiology*, **167**(1), 25–33. DOI: 10.1093/aje/kwm265.
- Klesges R., Debon M., Ray J. (1995) Are self-reports of smoking rate biased? Evidence from the Second National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology*, **48**(10), 1225–1233. DOI: 10.1016/0895-4356(95)00020-5.
- Komlos J. (1999) On the nature of the Malthusian threat in the eighteenth century. *Economic History Review*, **52**(4), 730–748. DOI: 10.1111/1468-0289.00145.
- Li Y., Wileyto E.P., Heitjan D.F. (2011) Statistical analysis of daily smoking status in smoking cessation clinical trials. *Addiction*, **106**(11), 2039–2046. DOI: 10.1111/j.1360-0443.2011.03519.x.
- Matt G., Jo Wilson S. (1994) Describing the frequency of marijuana use: Fuzziness and context-dependent interpretation of frequency expressions. *Evaluation and Program Planning*, **17**(4), 357–369. DOI: 10.1016/0149-7189(94)90036-1.
- Min Y., Agresti A. (2005) Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**(1), 1–19. DOI: 10.1191/1471082X05st084oa.
- Nuerk H.-C., Iversen W., Willmes K. (2004) Notational modulation of the SNARC and the MARC (linguistic markedness of response codes) effect. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, **57**(5), 835–863. DOI: 10.1080/02724980343000512.
- R Development Core Team. (2012) *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing.
- Rubin D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.
- Schneeweiss H., Komlos J., Ahmad A.S. (2010) Symmetric and asymmetric rounding: a review and some new results. *ASta Advances in Statistical Analysis*, **94**(3), 247–271. DOI: 10.1007/s10182-010-0125-2.
- Shiffman S. (2009a) Ecological momentary assessment (EMA) in studies of substance use. *Psychological Assessment*, **21**(4), 486–497. DOI: 10.1037/a0017074.
- Shiffman S. (2009b) How many cigarettes did you smoke? Assessing cigarette consumption by global report, timeline followback, and ecological momentary assessment. *Health Psychology*, **28**(5), 519–526. DOI: 10.1037/a0015197.
- Shiffman S., Gwaltney M.H., Balabanis M., Liu K., Paty J., Kassel J.D., Hickcox M., Gnys M. (2002) Immediate antecedents of cigarette smoking: an analysis from ecological momentary assessment. *Journal of Abnormal Psychology*, **111**(4), 531–545. DOI: 10.1037/0021-843X.111.4.531.
- Sobell L.C., Sobell M.B. (1992) Timeline followback: a technique for assessing self-reported alcohol consumption. In Litten R.Z., Allen J.P. (eds) *Measuring Alcohol Consumption: Psychosocial and Biochemical Methods*, pp. 41–72, Totowa, NJ: Humana Press.
- Wang H., Heitjan D.F. (2008) Modeling heaping in self-reported cigarette counts. *Statistics in Medicine*, **27**(19), 3789–3804. DOI: 10.1002/sim.3281.

- Wang H., Shiffman S., Griffith S.D., Heitjan D.F. (2012) Truth and memory: linking instantaneous and retrospective self-reported cigarette consumption. *The Annals of Applied Statistics*, **6**(4), 1689–1706. DOI: 10.1214/12-AOAS557.
- Wiederman M. (1997) The truth must be in here somewhere: examining the gender discrepancy in self-reported lifetime number of sex partners. *Journal of Sex Research*, **34**(4), 375–386. DOI: 10.1080/00224499709551905.
- Wileyto E.P., Patterson F., Niaura R., Epstein L.H., Brown R.A., Audrain-McGovern J., Hawk L. W., Lerman C. (2005) Recurrent event analysis of lapse and recovery in a smoking cessation clinical trial using bupropion. *Nicotine & Tobacco Research*, **7**(2), 257–268. DOI: 10.1080/14622200500055673
- Wutich A. (2009) Estimating household water use: a comparison of diary, prompted recall, and free recall methods. *Field Methods*, **21**(1), 49–68. DOI: 10.1177/1525822X08325673