Submitted to the Annals of Applied Statistics

HEAPING AND SEEPING, GAITD REGRESSION AND DOUBLY CONSTRAINED REDUCED-RANK VECTOR GENERALIZED LINEAR MODELS, IN SMOKING STUDIES

BY THOMAS YEE[1,*] AND LUCA FRIGAU[2] AND CHENCHEN MA [3]

[1]*Department of Statistics, University of Auckland,* [*]*t.yee@auckland.ac.nz*

[2]*Department of Economics and Business Sciences University of Cagliari, frigau@unica.it*

[3]*School of Mathematical Sciences and Center for Statistical Science, Peking University, chenchenma@pku.edu.cn*

Large-scale health surveys suitable for addiction studies furnish self-reported data that consequently suffer from a form of measurement error called heaping which statisticians have been grappling with for decades. Also known as digit preference, the aberration is often characterized by spikes at multiples of 10 or 5 upon rounding. To date, methods and software for heaped (and seeped) data have been largely wanting. Identifying three generic problems for simple addiction studies, we solve them by a newly developed technique called Generally Altered, Inflated, Truncated and Deflated (GAITD) regression for counts applied to a recent National Health and Nutrition Examination Survey data set. In conjunction, we propose the class of *Doubly constrained* Reduced-Rank VGLMs whereby the reduced-rank regression is afforded structure by way of linear constraints—this allows further simplification of the dimension reduction. We determine the distribution of smoking initiation age (SIA) and its association with tobacco consumption and smoking duration, e.g., is a lower SIA associated with higher tobacco consumption later in life? Is higher SIA associated with shorter smoking duration among quitters? Together, GAITD regression and DRR-VGLMs hold promise for heaped and seeped data.

1. Introduction. Addiction studies have flourished over the past century (Babor, 2000), as evidenced by outlets such as *Addiction* (since 1903), *Journal of Studies on Alcohol and Drugs* (1940), *Alcohol and Alcoholism* (1967), *Nicotine & Tobacco Research* (1999), as well as *Addiction Science & Clinical Practice* (2006). Addiction has been informally defined as "a process whereby a behavior, that can function both to produce pleasure and to provide relief from internal discomfort, is employed in a pattern characterized by (1) recurrent failure to control the behavior and (2) continuation of the behavior despite significant negative consequences" (Goodman, 1990). There are two broad addiction types: substance/drug and non-substance/behavioral (Zou et al., 2017). Although there are clear-cut diagnostic criteria for drug addiction set in the International Classification of Diseases and Related Health Problems (ICD-11, 2019, see also DSM-5-TR (2022)), a universally accepted formal definition for addiction is unlikely (Sussman and Sussman, 2011). During this century, the subject has broadened to include the internet, cellphone, vaping and fentanyl as specific examples.

While addiction science has undergone modernization in recent years through genetic screening (Buckland, 2008), Bayesian methods (West, 2016), machine learning (Mak, Lee and Park, 2019) etc., one remaining fundamental characteristic of all large-scale health studies collecting addiction data is self-reporting. Prior to devices to automatically monitor the behaviour of interest (which are essentially restricted to small-scale studies) such as wearable sensors (Mahmud et al., 2019), addiction data were mainly of the survey-type and therefore

Keywords and phrases: Addiction science, Count responses, Finite mixture distribution, Generally altered, inflated, truncated and deflated regression, Multinomial logit model, Reduced rank regression.

TABLE 1

Generic problems (P1)–(P3) ordered by time pertaining to a basic addiction study and phrased for smoking studies. (P4) is subsidiary to (P3). The primary response variables are $Y_{\rm SIA}$, $Y_{\rm TC}$, $Y_{\rm SD}$ where $Y_{\rm SD} = Y_{\rm SCA} - Y_{\rm SIA}$. With heaped and seeped counts, the problems are solved by GAITD regression. DRR-VGLMs are also developed and used.

Problem	Response	Set of problems and questions
(P1)	$Y_{\rm SIA}$	*Smoking Initiation Age*. What distribution does this have? Does it comprise of a small number of subgroups? If so, is there a high-risk subgroup? What covariates are associated with the distribution of $Y_{\rm SIA}$?
(P2)	$Y_{\rm TC}$	*Tobacco Consumption*. What distribution does this have? What covariates are associated with its distribution?—in particular, does it include SIA?
(P3)	$Y_{\rm SD}$	*Smoking Duration*. What distribution does this have? What covariates are associated with its distribution?—in particular, does it include SIA?
(P4)	$Y_{\rm SCA}$	*Smoking Cessation Age*. What distribution does this have?

suffered from a form of measurement error called heaping, also known as 'digit preference' or 'bunching'. Heaping commonly occurs in retrospective survey data. The aberration is frequently seen in an excess of multiples of 5 and 10 relative to other values. A classical example is the number of cigarettes smoked daily. Like food frequency questionnaires, it is uncommon for people to know their exact values and so recorded values suffer from measurement error (Carroll et al., 2006). For this question, often "1 pack" and "half a pack" are solicited which equates to 20 and 10 because in some countries such as USA a pack always amounts to 20 cigarettes (e.g., Figure 4a). Three examples of heaping can be seen in Figure 1 in a New Zealand cross-sectional study from the mid-1990s which recorded *tobacco consumption* ($Y_{\rm TC}$), *smoking duration* ($Y_{\rm SD}$) and *smoking cessation age* ($Y_{\rm SCA}$) in 5492 quitters. For instance, plot (b) shows 30 is heaped while 29 and 31 are seeped (too few observations). Current statistical methodology and software for analyzing heaping is inadequate and inaccessible (Yee and Frigau, 2025), however, a new technique called *Generally Altered, Inflated, Truncated and Deflated* (GAITD) regression (Yee and Ma, 2024) holds promise for such and is used as the central tool here.

In this paper our focus is on self-reported smoking studies that have count responses. Table 1 summarizes three generic problems applicable to simple addiction studies, abbreviated by (P1)–(P3). Phrased specifically for a smoking study, it is 'simple' because quitters are assumed not to relapse and are independent. We take the *smoking initiation age* ($Y_{\rm SIA}$) to be how old an individual is in years when the habit becomes practiced regularly, and the SCA to be when the person quits permanently. In between, TC is measured most commonly by the number of cigarettes smoked daily. These define $Y_{\rm SIA}$, $Y_{\rm SCA}$ and $Y_{\rm TC}$ respectively as three different response variables. A better alternative to $Y_{\rm SCA}$ is smoking duration $Y_{\rm SD} = Y_{\rm SCA} - Y_{\rm SIA}$ because its support is $\{0, 1, \ldots\}$ as an ordinary count distribution whereas $\mathrm{E}(Y_{\rm SCA}|Y_{\rm SIA})$ creates difficulties due to the constraint that $Y_{\rm SCA} < Y_{\rm SIA}$ under a generalized linear model-like framework.

Since our approach is regression, the distributions of the Y_j given covariates are of interest, and consequently (P2)–(P3) may be modelled by regressing $Y_{\rm TC}$ and $Y_{\rm SD}$ against $Y_{\rm SIA}$ because of possible temporal effects, e.g., it has been conjectured that the earlier a habit is formed the bigger the problem is later. For instance, Ali et al. (2020) showed that those who started smoking regularly at ages 18–20 years were more likely to experience high levels of nicotine dependence and less likely to attempt or intend to quit in adulthood compared with those who started at age 21 or older.

1.1. *On three generic problems in addiction studies.* Table 1 can be readily rephrased for other substance-based studies such as alcohol, a specific hard drug, or a behavioural addiction such as gambling. In our case, (P1)–(P3) concern ever-smokers, current and ex-smokers

respectively. Ever-smokers are those who have ever smoked. In the basic setting adopted, the (P2) and (P3) subsamples form a partition of the (P1) subsample, and the entire sample includes never-smokers.

Solving (P1)–(P3) has direct relevance for prevention, therapy and policy. For example,

(P1) Any high-risk subgroup that can be identified can be targeted for smoking prevention education, e.g., Khuder, Dayal and Mutgi (1999) expressed the need for prevention programs aimed at children below 16 years based on study participants who started smoking before that age having an odds ratio of 2.1 (CI 1.4–3.0) for not quitting smoking compared to those who started later than 16.

(P2) Econometricians have developed models such as the double-hurdle with the intention of determining the effects of economic factors, such as prices and income, and changing variables such as raising smoking tax, on personal tobacco consumption (e.g., Jones, 1989; Harris and Zhao, 2007) for use in policy making. If early SIA is associated with high TC then additional efforts should be made to postpone the beginning of smoking among youngsters, e.g., as recommended by Taioli and Wynder (1991) who by comparing ≤ 14 versus ≥ 20 year olds showed an early age smoking habit led to a heavy cigarette consumption later in life. In theory, GAITD regression allows a different set of explanatory variables to be modelled by each special value type. Thus, for example, variable selection can be used to identify covariates important for discriminating between heavy versus light smokers and this is potentially helpful for subgroup-specific intervention.

(P3) and **(P4)** Covariates associated with quitting may be exploited by smoking cessation and reduction therapies, e.g., we show in Section 5.2 that exposure to passive smoke is positively associated with a higher TC, hence potential quitters should shun such environments or consider cessation programs for couples if married (Choi, 2022). Two examples of covariates found in smoking studies relating to successful smoking cessation include Qiu et al. (2020) who showed success was associated with older age, self-perceived poor health and fewer cigarettes smoked daily, and Eum et al. (2022) who showed that quitting was influenced by the presence of hypertension or a cardiovascular disease. If SIA is a significant covariate for SCA then there are various consequences, for example, (i) quitting at younger ages give larger reductions in excess mortality associated with continued smoking (Thomson et al., 2022); (ii) the making of public health policy, for instance the T21 policy enacted in USA (`tobacco21.org`; Ribisl and Mills, 2019; Ali et al., 2020) to raise the sales age for all nicotine and tobacco products to 21, and New Zealand's 2023 ban on cigarette sales to those under 18 years of age and tobacco denicotinisation (Wilson et al., 2022). Knowing the cessation age can also aid practical access to subgroups, e.g., 65–70 year olds in social clubs can be canvassed.

To summarize, given count responses Y_{SIA}, Y_{TC}, Y_{SD} that are contaminated by heaping and seeping, the main purpose of this paper is to solve (P1)–(P4) for a large recent US smoking data set described in Section 1.4 by GAITD regression and a new technique abbreviated by "DRR-VGLMs" proposed in Section 4. Although GAITD regression is usually fitted as a vector generalized linear model (VGLM), a reduced-rank VGLM (RR-VGLM) is a useful variant, and doubly constrained RR-VGLM (DRR-VGLM) is an enhanced extension of RR-VGLMs developed in this paper.

1.2. *On heaped and seeped data.* Heaped data often results from rounding responses retrospectively when the precise values are unknown due to mis-remembering. Other causes include retrieval failure (e.g., Krinsky and Nelson, 1985), satisficing (e.g., Artinger, Gigerenzer and Jacobs, 2022) and intentional mis-reporting, Heaping can be defined (especially for

financial data) as *the tendency to provide estimates ending in digits that are the largest divisors of the base number system* (Jorgensen, Patrick and Soderstrom, 2020, p. 178). Consequently, the distorted distribution is characterized by spikes, especially at multiples of 5 or 10. Seeped data, by comparison, has received far less attention and some authors such as Wolff and Augustin (2003) refer to the phenomenon as *downward heaping*, with the usual case called *upward* heaping. Seeped data are characterized by dips (or holes) due to a lack of data compared to what is nominally expected. Often the immediate values adjacent to a heaping point are seeped, because the increase and decrease in probabilities balance out to some extent, e.g., years 29 and 31 in Figure 1(c) are seeped. As heaped data is very common it is of no surprise that it has been encountered in many different fields. Yee and Frigau (2025) reviews heaped and seeped data from many application areas, as well as methods and software that have been proposed for handling such.

1.3. *Smoking studies and heaping.* Since Doll and Hill (1950), smoking has justifiably become implicated as a risk-factor for many diseases and medical conditions. Its position as a cause for harm is now undeniable, and smoking studies have fueled much of their impetus from this. It has now grown into a large field with active subfields such as cigarette consumption, smoking cessation and community prevention measures (Brown et al., 1998; Wang and Heitjan, 2008; Bar and Lillard, 2012; Jung, Choi and Park, 2018) which can lead to associated tax and political issues (Forster and Jones, 2001). Statisticians have actively contributed to the field (e.g., Klesges, Debon and Ray, 1995; Lewis-Esquerre et al., 2005; Wang and Heitjan, 2008; Wang et al., 2012; Allen et al., 2017) by grappling with the heaping problem. Here, smokers often use 5 and its multiples to report daily cigarette consumption. The empirical distribution is characterized by the peak being a multiple of 5 (Cummings et al., 2015; Jung, Choi and Park, 2018) or of 10 (Klesges, Debon and Ray, 1995). The latter observed that social-demographic characteristics influence self-reported daily cigarette consumption, and upon interviewing $20,322$ smokers, they found that heavier smokers, Caucasians, and those with less education were more likely to exhibit a digit preference than lighter smokers, African-Americans and those with more education. Generally, smokers prefer to report consumption equal to some multiple of 10 cigarettes with the most common response being 20 daily. Some of these observations can be seen *spikeplotted* (Cox, 2004) in Figures 1 and 4.

Clearly, heaping and seeping must be adjusted for in a statistical analysis because in theory a spike could almost be arbitrarily large relative to the main distribution, and located at any position, e.g., Christelis and Sanz-de Galdeano (2009) found that accounting for heaping in smoking studies improved the fit of (regression) models considerably. Our analyses show this to be true too, e.g., Fig. 4. This is confirmed in our analyses by comparisons with naive negative binomial regressions. Despite Crawford, Weiss and Suchard (2015, p. 572) stating that "inference for heaped data is an important statistical problem", Yee and Frigau (2025) shows that to date, statistical methods for such have been largely sparse and inadequate. Fortunately, a newly developed method called GAITD regression has been proposed (Yee and Ma, 2024) and implemented in software. Coupled with DRR-VGLMs, heaped and seeped data can potentially be analyzed by an exceedingly flexible regression model that allows a simplified reduced-rank regression.

1.4. *The NHANES 2017–2020 data.* Smoking and covariate data were extracted from the combined 2017–2018/2019–2020 cycles of the National Health and Nutrition Examination Survey (NHANES; Zipf et al., 2013). The second cycle was never completed because of Covid-19 so all data just prior to the pandemic were merged with the 2017–2018 cycle to produce a larger "`P_`"-type data set. From this, the data frame `smqP` is available in the VGAMdata R package (Yee and Gray, 2025). Most smoking studies adjust for age, marital

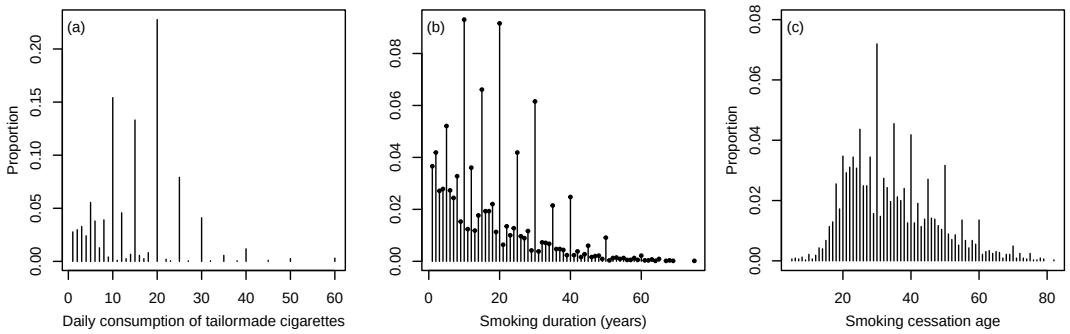

FIG 1. *Three spikeplotted examples of heaping from a cross-sectional study* `xs.nz` *in* VGAMdata *involving 5492 current or past smokers. (a) Tobacco consumption (Y_{TC}); (b) Smoking duration (Y_{SD}); (c) Age at which the past smokers quit (Y_{SCA}).*

status, education and race, and following suit it was necessary to delete cases due to missing values in these variables. Furthermore, age and SCA were right-censored. This necessitated further deletions to simplify the overall interpretation. Because we restricted our analyses to those no older than 79 and who had quit less than 50 years ago, our interpretation is conditional on these range restrictions. After processing, there remained 3327 participants with a smoking history, comprising 1750 ex-smokers and 1577 current smokers. Race was simplified from 5 levels to binary (white non-Hispanics versus others), and total exposure to secondary smoke was computed by summing its presence/absence over seven sources, such as restaurants and bars, so that totals ranged from 0 to 7.

There is clear evidence of overdispersion relative to a Poisson distribution for all the responses, e.g., the variance-to-mean ratio for SCA is 4.79, therefore we adopted the negative binomial distribution (NBD; (4)) as the main distribution throughout, called the parent f_π in (1) below. Another reason is that our responses are usually unimodal, hence strictly monotonic distributions such as the zeta (Hu et al., 2006) and logarithmic were deemed unsuitable.

It is noted that SCA is computed as the difference between current age and quitting duration. Because it was not measured directly, it is only weakly heaped (Figure 8a). In contrast, SCA is recorded directly in some smoking studies such as in MacMahon et al. (1995) whose Figure 1(a)–(b) is strongly heaped. Clearly, the amount of heaping depends on how direct questions are posed to study participants.

2. The GAITD 'combo' model.

2.1. *Notation and the Operators.* Let $\mathcal{S} = \{\mathcal{A}_p, \mathcal{A}_{np}, \mathcal{I}_p, \mathcal{I}_{np}, \mathcal{T}, \mathcal{D}_p, \mathcal{D}_{np}\}$ be the set of *special values* in the support of the parent (or base) distribution f_π—these values are directly affected by alteration, inflation, truncation or deflation. All the subsets are mutually exclusive. Subscripts p and np are for 'parametric' and 'nonparametric'. The GAITD 'combo' is a super mixture model involving f_π which is possibly modified by f_α, f_ι and f_δ having generic parameters $\boldsymbol{\theta}_\alpha$, $\boldsymbol{\theta}_\iota$ and $\boldsymbol{\theta}_\delta$. These distributions describe the parametric parts defined on $\mathcal{A}_p$, $\mathcal{I}_p$, $\mathcal{D}_p$. The nonparametric parts are modelled by patternless probabilities described in Section 3. Subscripts π, α, ι and δ are thus used to distinguish between the (parametric) *p*arent, *a*ltered, *i*nflated and *d*eflated distributions respectively, e.g., the zero-inflated Poisson is a $\mathrm{Pois}(\mu_\pi) + \{0\}$ mixture written in our notation as a GI–Pois($\theta_\pi = \mu_\pi$) with $\mathcal{I}_{np} = \{0\}$, because "GI" means general inflation only. The elements of $\mathcal{A}_p$ and $\mathcal{A}_{np}$ are written $a_1, a_2, \ldots$ with the weakness that they are not distinguishable being irrelevant here. Likewise $i_1, i_2, \ldots$, and $t_1, t_2, \ldots$, and $d_1, d_2, \ldots$ for the other operators.

The probabilities used generally for alteration, inflation and deflation are ω, ϕ and ψ respectively. Special values are enumerated in two ways, e.g., if 20 is the only heaped value handled by inflation then the spike probability is ϕ_1 or $\phi_{\lceil 20 \rceil}$. We let $|\mathcal{A}|$ denote the cardinality of a set $\mathcal{A}$, and $\mathcal{R}$ the support of a distribution. Concerning matrices and vectors, $\mathbf{A}_{ij} = a_{ij}$, $\mathbf{0}_n$ and $\mathbf{1}_n$ are n-vectors of 0s and 1s, with $\boldsymbol{e}_k$ being $\mathbf{0}_n$ but with a 1 in the kth position. Superscript T denotes transpose.

A short summary of the four fundamental operators (in decreasing order of precedence) is as follows. Some specimen questions are given concerning $Y = Y_{\mathrm{TC}}$ (in cigarettes per day).

- *Truncation*. $\Pr(Y = y) = 0$ for some $y \in \mathcal{T}$. Certain values of y are impossible. For example, among current smokers, TC has $\mathcal{T} = \{0\}$.
- *Alteration*. $\Pr(Y = y)$ is modelled separately from $\Pr(Y \neq y)$. Handles heaped and seeped values. Answers 'why are observations there?' e.g., what is the probability a smoker consumes one pack daily? Answer: $\omega_{\lceil 20 \rceil}$. An example from Figure 4 are those who smoke half or one pack daily—then $\omega_{\lceil 10 \rceil} + \omega_{\lceil 20 \rceil} \approx \frac{1}{3}$.
- *Inflation*. $\Pr(Y = y)$ is a sum from two sources and handles heaping. Answers 'why are observations there *in excess*?' e.g., in Figure 1(a), $\phi_{\lceil 10 \rceil}$ is the spike probability rising above the ordinary TC distribution and is obviously due to heaping. The same feature concerning half a pack can be seen in Figure 4.
- *Deflation*. The opposite of inflation, $\Pr(Y = y)$ handles seeping. Answers 'why are observations *not there*?' e.g., in Figure 1(a), if $\psi_{\lceil 9 \rceil} + \psi_{\lceil 11 \rceil} = \phi_{\lceil 10 \rceil}$ then seeping at 9 and 11 could explain the heaping at 10.

2.2. *Probability Mass Function.* With a 1-parameter parent, the GAITD 'combo' probability mass function (PMF) is $\Pr(Y = y; \theta_\pi, \omega_p, \theta_\alpha, \phi_p, \theta_\iota, \psi_p, \theta_\delta, \boldsymbol{\omega}_{np}, \boldsymbol{\phi}_{np}, \boldsymbol{\psi}_{np}) =$

$$
(1)\quad \begin{cases} 0, & y \in \mathcal{T}, \\ \omega_p \, f_\alpha(y) \,/ \sum\limits_{u \in \mathcal{A}_p} f_\alpha(u), & y \in \mathcal{A}_p, \\ \omega_s, & y = a_s \in \mathcal{A}_{np},\ s = 1, \ldots, |\mathcal{A}_{np}|, \\ \Delta \, f_\pi(y) + \phi_p \, f_\iota(y) \,/ \sum\limits_{u \in \mathcal{I}_p} f_\iota(u), & y \in \mathcal{I}_p, \\ \Delta \, f_\pi(y) + \phi_s, & y = i_s \in \mathcal{I}_{np},\ s = 1, \ldots, |\mathcal{I}_{np}|, \\ \Delta \, f_\pi(y) - \psi_p \, f_\delta(y) \,/ \sum\limits_{u \in \mathcal{D}_p} f_\delta(u), & y \in \mathcal{D}_p, \\ \Delta \, f_\pi(y) - \psi_s, & y = d_s \in \mathcal{D}_{np},\ s = 1, \ldots, |\mathcal{D}_{np}|, \\ \Delta \, f_\pi(y), & y \in \mathcal{R} \setminus \{\mathcal{A}_p,\ \mathcal{A}_{np},\ \mathcal{I}_p,\ \mathcal{I}_{np},\ \mathcal{T},\ \mathcal{D}_p,\ \mathcal{D}_{np}\}, \end{cases}
$$

where the normalizing constant $\Delta =$

$$
(2)\quad \left(1 - \omega_p - \phi_p + \psi_p - \sum_{u=1}^{|\mathcal{A}_{np}|} \omega_u - \sum_{u=1}^{|\mathcal{I}_{np}|} \phi_u + \sum_{u=1}^{|\mathcal{D}_{np}|} \psi_u\right) \cdot \left[1 - \sum_{a \in \mathcal{A}} f_\pi(a) - \sum_{t \in \mathcal{T}} f_\pi(t)\right]^{-1}.
$$

A Shiny app (Chang et al., 2023) at `www.stat.auckland.ac.nz/~yee` enables the PMF to be explored and a data set to be uploaded for model building and initial values.

The overall structure of (1) is simple: after truncating certain values and specifying certain altered values and its probabilities, $\Delta \, f_\pi(y)$ is the scaled parent from which spikes can be added or dips subtracted to allow for inflation and deflation respectively. The spikes or heaped values may be unstructured (ϕ_s) or come from another distribution with PMF f_ι on the set of values from $\mathcal{I}_p$. Likewise, the dips or seeped values may be unstructured (ψ_s) or come from another distribution f_δ on $\mathcal{D}_p$. Deflation is therefore the opposite of inflation because of the subtraction versus the addition. Offering greater flexibility, one can 'combine' inflation and

deflation by alteration so that spikes and dips can be accommodated by ω_s on $\mathcal{A}_{np}$ and $\omega_p f_\alpha$ on $\mathcal{A}_p$. The bottom equation of (1) is for non-special values that receive no special treatment apart from the scaling needed to adjust for the special values.

Thus there are two types of each of alteration, inflation and deflation: parametric and nonparametric. The latter specify probabilities that are unstructured, e.g., any values that are inexplicable, are outliers or are nuisance, would more aptly be handled nonparametrically. The values are taken into account but they usually contribute less insight to the overall model because they do not convey any information about $\boldsymbol{\theta}_\pi$, $\boldsymbol{\theta}_\alpha$, $\boldsymbol{\theta}_\iota$ or $\boldsymbol{\theta}_\delta$. In contrast, the parametric variants utilize the offspring distributions f_α, f_ι and f_δ, and for convenience they are the same as f_π with possibly different parameter values. If the mechanism generating the heaping is the same as f_π but at a greater sampling intensity, then parametric inflation is suitable. Totally parametric GAITD regression is appropriate if the heaping and seeping is affecting the entire distribution equally, cf. non-differential measurement error.

Figure 2 illustrates (1) by displaying a hypothetical smoking cessation age distribution based on Fig. 1 afflicted by heaping/seeping. The parent is the NBD. Plot (c) resembles the ideal situation where spikes can be explained by adjacent dips, and plot (d) conveys how some values may be inexplicably deviant from the main distribution. Figure 2(a) mimics heaping where the heap values are sampled from another NBD but with greater intensity—this is parametric inflation. Figure 2(b) is parametrically generally-altered: the special values are also NB distributed but have less probability than usual, hence dips are created. Figure 2(c) is the combination of (a) and (b). Figure 2(d) is nonparametrically generally-altered because the probabilities at $\mathcal{A}_{np}$ are not NB distributed—they are patternless and different from the main distribution.

Because (1) is a mixture, identifiability constraints are necessary for estimability, e.g., where applicable, $|\boldsymbol{\theta}_\alpha| < |\mathcal{A}_p|$ and $|\mathcal{I}_p| \neq 1$ by convention. Readers are directed to the list of conditions given in Yee and Ma (2024, sec. 5.3) for details. Like other ill-posed models, problems associated with identifiability can often be detected by monitoring convergence during estimation. In our Section 5 analyses we have attempted to fit well-conditioned models that avoid common problems which might beset novices.

3. VGLMs, VGAMs and the multinomial logit model. In a nutshell, GAITD regression fits (1) as a vector generalized linear model involving a multinomial logit model (MLM), so it is expedient to summarize them now. Like GLMs, the covariate–response data is assumed to be $(\boldsymbol{x}_i,\ y_i)$, $i = 1, \ldots, n$, independently where $\dim(\boldsymbol{x}_i) = d$ with $x_{i1} = x_1 = 1$ denoting the intercept for the ith individual and first explanatory variable respectively. The log-likelihood, such as based on (1), is $\ell = \sum_{i=1}^n w_i^* \ell_i$ where allowance is made for positive, known and prespecified prior weights w_i^*. VGLMs apply linear predictors η_j (j indexes the jth linear predictor) to model multiple parameters so that usually

$$g_j(\theta_j) \ = \ \eta_j \ = \ \boldsymbol{\beta}_j^T \boldsymbol{x} \ = \ \sum_{k=1}^{d} \beta_{(j)k}\, x_k, \quad j = 1, \ldots, M, \tag{3}$$

(where $M = \dim((\eta_1, \ldots, \eta_M)^T) = \dim(\boldsymbol{\eta})$ by definition) for some suitable (parameter) link function g_j applied to θ_j, however, the MLM link involves multiple η_j. For example, since our parent is the NBD of R's `dnbinom()`, then

$$\eta_1 \ = \ \log \mu_\pi, \qquad \eta_2 \ = \ \log \kappa_\pi, \tag{4}$$

where $\boldsymbol{\theta}_\pi = (\mu_\pi,\ \kappa_\pi)^T$ comprise the positive mean and index parameters (Lawless, 1987; Hilbe, 2011; Yee, 2020) with $\text{Var}(Y) = \mu_\pi[1 + \mu_\pi/\kappa_\pi]$.

As $M > 1$, linear constraints on the regression coefficients are accommodated by

$$\boldsymbol{\eta}(\boldsymbol{x}_i) \;=\; \begin{pmatrix} \eta_1(\boldsymbol{x}_i) \\ \vdots \\ \eta_M(\boldsymbol{x}_i) \end{pmatrix} \;=\; \sum_{k=1}^{d} \boldsymbol{\beta}_{(k)}\, x_{ik} \;=\; \sum_{k=1}^{d} \mathbf{H}_k\, \boldsymbol{\beta}^*_{(k)}\, x_{ik} \;=\; \mathbf{B}^T \boldsymbol{x}_i, \tag{5}$$

for known *constraint matrices* $\mathbf{H}_k$ of full column-rank, and $\boldsymbol{\beta}^*_{(k)} = (\beta^*_{(1)k}, \beta^*_{(2)k}, \ldots)^T$ is a possibly reduced set of regression coefficients to be estimated. While trivial constraints are denoted by $\mathbf{H}_k = \mathbf{I}_M$, other common examples include parallelism ($\mathbf{H}_k = \mathbf{1}_M$), and *intercept-only* parameters $\eta_j = \beta^*_{(j)1}$ where all $\beta_{(j)k}$ but the intercept are set to 0. For example, the NBD (4) with $d = 3$ explanatory variables and intercept-only κ_π would have $\mathbf{H}_1 = \mathbf{I}_2$ and $\mathbf{H}_2 = \mathbf{H}_3 = (1,\ 0)^T$ so that $\log \mu_\pi = \beta^*_{(1)1} + \beta^*_{(1)2}\, x_2 + \beta^*_{(1)3}\, x_3$ and $\log \kappa_\pi = \beta^*_{(2)1}$. Later, we utilize the parallelism assumption for certain x_k a number of times in our analyses to simplify their effect to a single regression coefficient. For DRR-VGLMs, the $\mathbf{H}_k$ are equal and estimated, and have constraint matrices to describe its form.

Like GLMs, VGLMs are estimated by the iteratively reweighted least squares (IRLS) algorithm which produces the variance-covariance matrix $\widehat{\text{Var}}\left(\widehat{\boldsymbol{\beta}^*}\right) = \left(\mathbf{X}_{\text{VLM}}^T\, \widehat{\mathbf{W}}\, \mathbf{X}_{\text{VLM}}\right)^{-1}$ evaluated at the final iteration, where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^{*T}_{(1)}, \ldots, \boldsymbol{\beta}^{*T}_{(d)})^T$ are all the regression coefficients to be estimated, $\mathbf{X}_{\text{VLM}}$ is the 'large' model matrix, and $\mathbf{W}$ is the overall working weight matrix, so that $\widehat{\boldsymbol{\beta}^*}$ is the MLE. The square roots of the diagonal elements of $\widehat{\text{Var}}\left(\widehat{\boldsymbol{\beta}^*}\right)$ are used for the standard errors in Tables 2–4.

Suitable for exploratory data analysis, vector generalized additive models (VGAMs) generalize VGLMs so that each η_j is an additive predictor like an ordinary GAM (e.g., Hastie and Tibshirani, 1990; Wood, 2017). They relax the linearity assumption of (5) to

$$\boldsymbol{\eta}(\boldsymbol{x}_i) \;=\; \sum_{k=1}^{d} \mathbf{H}_k\, \boldsymbol{f}^*_{(k)}(x_{ik}), \tag{6}$$

where the $\boldsymbol{f}^*_{(k)}(x_{ik}) = (f^*_{(1)k}(x_{ik}), \ldots, f^*_{(R_k)k}(x_{ik}))^T$ are vectors of smooth component functions estimated by vector splines (Fessler, 1991) and backfitting. More details may be found in Yee and Wild (1996) and Yee (2015, ch. 4).

The multinomial logit model generalizes the logit link to more than a single probability. Given a vector of probabilities $\boldsymbol{p} = (p_1, \ldots, p_D)^T$ say, $g = \text{multilogit}(p_1, \ldots, p_D)$ is

$$g(p_s) \;=\; \eta_s = \log\left\{p_s / p_{D+1}\right\}, \qquad s = 1, \ldots, D,$$

where $p_{D+1} = 1 - \sum_{u=1}^{D} p_u$ corresponds to the baseline group. The inverse link (also called the softmax function) is $p_s = e^{\eta_s} / \sum_{u=1}^{D+1} e^{\eta_u}$ where $\eta_{D+1} \equiv 0$ for identifiability.

The entire model fitted as a VGLM (5) has $\boldsymbol{\eta}^T = (\eta_1, \ldots, \eta_M)^T =$

$$\left(g_\pi(\theta_\pi),\ \log\frac{\omega_p}{\mathcal{N}},\ g_\alpha(\theta_\alpha),\ \log\frac{\phi_p}{\mathcal{N}},\ g_\iota(\theta_\iota),\ \log\frac{\psi_p}{\mathcal{N}},\ g_\delta(\theta_\delta),\ \log\frac{\omega_1}{\mathcal{N}},\ \ldots,\ \log\frac{\omega_{|\mathcal{A}_{np}|}}{\mathcal{N}}, \right.$$

$$\left. \log\frac{\phi_1}{\mathcal{N}},\ \ldots,\ \log\frac{\phi_{|\mathcal{I}_{np}|}}{\mathcal{N}},\ \log\frac{\psi_1}{\mathcal{N}},\ \ldots,\ \log\frac{\psi_{|\mathcal{D}_{np}|}}{\mathcal{N}} \right) \tag{7}$$

where $g_\cdot(\cdot)$ are the link functions. The quantity $\mathcal{N} = 1 - \omega_p - \phi_p - \psi_p - \sum\omega_u - \sum\phi_u - \sum\psi_u$ corresponds to the reference group. Thus in summary, GAITD regression ties together (1) and (7). In the analyses, Tables 2–4 have regression coefficients ordered by (7).

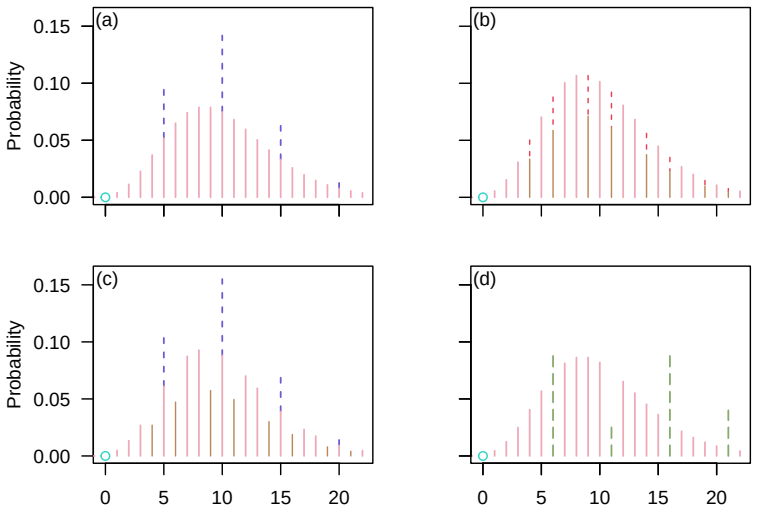

FIG 2. *Heaped and/or seeped data—some idealized forms based on a NBD parent. Plots (a)–(c) are parametric GAITD–NB PMF and (d) is a nonparametric GAITD–NB PMF. The special values are* $\mathcal{T} = \{0\}$, $\mathcal{I}_p = \{5, 10, 15, 20\}$, $\mathcal{D}_p = \{4, 6, 9, 11, 14, 16, 19, 21\}$ *with* $\phi_p = 0.15$, $\psi_p = 0.15$, $\mu_\pi = 10$, $k_\pi = 10$ *so that* $f_\pi = f_\iota = f_\delta$ *are the NB(10, 10) PMF (see (4)). (a) GIT–NB; (b) GTD–NB with the dip probabilities shown; (c) GITD–NB combines them together; (d) GAT–NB–MLM(*$\boldsymbol{\omega}_{np} = (0.09, 0.03, 0.09, 0.04)^T$*). In the electronic copy, the parent distribution comprises the solid pink lines, truncated values are hollow turquoise circles, the altered probabilities are dashed avocado lines, and the inflation probabilities are dashed indigo spikes. Also the deflation probabilities are solid brown (e.g., dirt or deer coloured) lines, and their differences are dashed red lines. A* Shiny *app for exploring (1) is available at* `www.stat.auckland.ac.nz/~yee`

Because terms in (7) involving $\mathcal{N}$ correspond to MLM probabilities, nonparametric alteration/inflation/deflation swiftly creates a complex model involving potentially many regression coefficients, especially when there are covariates. This provides strong motivation for proposing the DRR-VGLM class.

4. Doubly constrained RR-VGLMs. Reduced-rank regression has been found to be a successful dimension reduction method in a range of settings (e.g., Anderson, 1951; Izenman, 1975; Bura et al., 2018; Powers, Hastie and Tibshirani, 2018; Reinsel, Velu and Chen, 2022). When applied to the VGLM class, reduced-rank VGLMs (RR-VGLMs; Yee and Hastie, 2003) approximate a subset of $\mathbf{B}$ in (5) so that far fewer regression coefficients may need to be estimated. In general, $\mathbf{B}$ is $d \times M$, dense (not sparse) and of rank $\min(d, M)$, however for some data sets, both M and d are "too" large and result in overfitting, e.g., a multinomial logit model with $M + 1 \gg 2$ levels. For GAITD regression with covariates, having a large value for $|\mathcal{A}_{np}| + |\mathcal{I}_{np}| + |\mathcal{D}_{np}|$ produces many MLM coefficients because nonparametric special values are modelled individually. Consequently, DRR-VGLMs potentially allow a much simpler regularization, e.g., Sections 5.3.1 and 5.4.1.

A skeletal description of RR-VGLMs is as follows. Partition $\boldsymbol{x}$ into $(\boldsymbol{x}_{[1]}^T, \boldsymbol{x}_{[2]}^T)^T$ and $\mathbf{B} = (\mathbf{B}_{[1]}^T \, \mathbf{B}_{[2]}^T)^T$ accordingly. RR-VGLMs approximate (5) by

$$\boldsymbol{\eta} = \mathbf{B}_{[1]}^T \boldsymbol{x}_{[1]} + \mathbf{A}\,\mathbf{C}^T \boldsymbol{x}_{[2]} = \mathbf{B}_{[1]}^T \boldsymbol{x}_{[1]} + \mathbf{A}\,\boldsymbol{\nu} \quad \text{where} \tag{8}$$

$$\mathbf{C} = \left(\boldsymbol{c}_{(1)} \cdots \boldsymbol{c}_{(R)}\right) = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{d_2})^T \text{ is } d_2 \times R, \tag{9}$$

$$\mathbf{A} = \left(\boldsymbol{a}_{(1)} \cdots \boldsymbol{a}_{(R)}\right) = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M)^T \text{ is } M \times R, \tag{10}$$

with $d_1 + d_2 = d$. One may interpret $\boldsymbol{\nu}$ as a vector of R latent variables $\nu_r = \boldsymbol{c}_{(r)}^T \boldsymbol{x}_{[2]}$. When the *rank* $R \ll \min(d_2, M)$ then effectively $\mathbf{B}_{[2]}$ is the outer product of two thin/tall but completely general matrices. For identifiability, since

$$\mathbf{A}\mathbf{C}^T = \mathbf{A}\,\mathbf{M}^{-1}\,\mathbf{M}\,\mathbf{C}^T \tag{11}$$

for any nonsingular $\mathbf{M}$, (full) corner constraints $\mathbf{A} = (\mathbf{I}_R, \widetilde{\mathbf{A}}^T)^T$ may be used. This feature follows 'corner point constraints' in the ANOVA literature. The elements of $\mathbf{B}_{[1]}$, $\widetilde{\mathbf{A}}$ and $\mathbf{C}$ are unconstrained and estimable.

We now propose *doubly constrained* RR-VGLMs (DRR-VGLMs) which offer a significant enhancement over RR-VGLMs by providing structure to $\mathbf{A}$ and $\mathbf{C}$. For example, suppose $\boldsymbol{x}_{[2]}^T = (\boldsymbol{x}_{[2]}^{\dagger T}, \boldsymbol{x}_{[2]}^{\ddagger T})$ where $\boldsymbol{x}_{[2]}^{\dagger}$ are physical variables and $\boldsymbol{x}_{[2]}^{\ddagger}$ are psychological variables from a health study. We might insist on two latent variables comprising linear combinations of each separately, i.e.,

$$\nu_1 \;=\; \boldsymbol{c}^{\dagger T} \boldsymbol{x}_{[2]}^{\dagger} \;\text{ and }\; \nu_2 \;=\; \boldsymbol{c}^{\ddagger T} \boldsymbol{x}_{[2]}^{\ddagger}. \tag{12}$$

Thus there is greater control in the reduced-rank regression, leading to a finer interpretation, e.g., they allow user-specified elements in $\mathbf{A}$ and $\mathbf{C}$ to be set to 0.

To allow such flexibility, each column of $\mathbf{A}$ and row of $\mathbf{C}$ is afforded its own constraint matrix, e.g., (12) may be fitted by $\mathbf{H}_{\mathrm{C}k}^{\dagger} = (1,\ 0)^T$ and $\mathbf{H}_{\mathrm{C}k}^{\ddagger} = (0,\ 1)^T$. Firstly for $\mathbf{C}$,

$$\mathbf{C}^T \;=\; \sum_{k=1}^{d_2} \boldsymbol{e}_k^T \otimes [\mathbf{H}_{\mathrm{C}k}\, \boldsymbol{c}_k^*]\,. \tag{13}$$

where $\otimes$ is the Kronecker product. Each user-specified $\mathbf{H}_{\mathrm{C}k}$ is $R \times R_{\mathrm{C}k}$ for $1 \le R_{\mathrm{C}k} \le R$ so the rows of $\mathbf{C}$ are regularized. Secondly for $\mathbf{A}$, *(restricted)* corner constraints (RCCs) are chosen so that

$$\mathbf{A} \;=\; \sum_{r=1}^{R} \boldsymbol{e}_r^T \otimes [\mathbf{H}_{\mathrm{A}r}\, \boldsymbol{a}_r] \;=\; \mathbf{A}_{\mathrm{RCC}} + \sum_{r=1}^{R} \boldsymbol{e}_r^T \otimes [\mathbf{H}_{\mathrm{A}r}^*\, \boldsymbol{a}_r^*] \;=\; \mathbf{A}_{\mathrm{RCC}} + \mathbf{A}^*, \tag{14}$$

where $\mathbf{A}_{\mathrm{RCC}}$ is a matrix of known constants and the $\mathbf{H}_{\mathrm{A}r}^*$ are a subset of the columns of $\mathbf{H}_{\mathrm{A}r}$. Combining (13) and (14), DRR-VGLMs are defined by

$$\boldsymbol{\eta} \;=\; \mathbf{B}_{[1]}^T \boldsymbol{x}_{[1]} + \left\{ \mathbf{A}_{\mathrm{RCC}} + \sum_{r=1}^{R} \boldsymbol{e}_r^T \otimes [\mathbf{H}_{\mathrm{A}r}^*\, \boldsymbol{a}_r^*] \right\} \left\{ \sum_{k=1}^{d_2} \boldsymbol{e}_k^T \otimes (\mathbf{H}_{\mathrm{C}k}\, \boldsymbol{c}_k^*) \right\} \boldsymbol{x}_{[2]} \tag{15}$$

so that $\mathbf{B}_{[1]}$ and the starred vectors are estimated. This may be done by an alternating sub-algorithm that accommodates constraint matrices, which is nested within the IRLS algorithm mentioned in Section 3. For this, (i) fix $\mathbf{C}$ and estimate $\mathbf{A}^*$ using the $\mathbf{H}_{\mathrm{A}r}^*$ as constraints; (ii) fix $\mathbf{A}$ and estimate $\mathbf{C}$ using $\mathbf{A}\mathbf{H}_{\mathrm{C}k}$ as constraints. Justification for the sub-algorithm can be obtained by a close examination of (5) where $\mathbf{A}$ takes on the role of $\mathbf{H}_k$ and $\mathbf{B}_{[2]}$ alternately. DRR-VGLMs may be applied to the 100+ regression models in VGAM.

It is noted that DRR-VGLMs are limited to *separable* problems (13)–(14) because otherwise existing structure in $\mathbf{A}$ and $\mathbf{C}$ may be destroyed by imposing full corner constraints (11). Let $\boldsymbol{i} = (i_1, \ldots, i_R)^T$ be the RCC indices, e.g., $\boldsymbol{i}^T = (1, \ldots, R)$ for ordinary RR-VGLMs. Separability means that, for $r = 1, \ldots, R$, the i_rth row of $\mathbf{H}_{\mathrm{A}r}$ has a single nonzero value and the i_rth row of $\mathbf{H}_{\mathrm{A}s}$ is $\mathbf{0}^T$ for all $s \ne r$. Hence we refer to (15) as having RCCs. Using R-style notation, the number of independent parameters in $\mathbf{A}$ is $\sum_{r=1}^{R} \texttt{ncol}(\mathbf{H}_{\mathrm{A}r}^*)$.

The overall covariance matrix is computed by constructing a block matrix comprising 6 unique blocks of which only $\mathrm{Var}(\mathrm{vec}(\widehat{\boldsymbol{a}}_r^*), \mathrm{vec}(\widehat{\boldsymbol{c}}_k^*))$ poses difficulties and the profile likelihood (Richards, 1961) is employed for this.

To justify the above, we let $\mathbf{M} = \text{diag}(a_{i_1 i_1}, \ldots, a_{i_R i_R})$ in (11) so that rows $\boldsymbol{i}$ of $\mathbf{A}\,\mathbf{M}^{-1}$ have the form of $\mathbf{I}_R$ but are subject to the $\mathbf{H}_{Ar}$. The off-diagonal elements are already zero because of the separability property. Thus the RCC transformation preserves constraints by carrying over any necessary values. As an example, if $a, b, \ldots$ are free parameters and $\boldsymbol{i} = (3,\ 4)^T$ for

$$\mathbf{A} = \begin{pmatrix} a & c \\ b & c \\ b & 0 \\ 0 & c \\ 0 & d \end{pmatrix} \quad \text{then } \mathbf{A}_{\text{RCC}} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and } \mathbf{A}^* = \begin{pmatrix} a^* & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & d^* \end{pmatrix}$$

is solved for the $4 - 2 = 2$ unconstrained asterisked elements because the theoretical and alternating-enabled constraint matrices are

$$\mathbf{H}_{\text{A1}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{H}_{\text{A2}} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{H}^*_{\text{A1}} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{H}^*_{\text{A2}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

The problem becomes nonseparable for $\boldsymbol{i} = (1,\ 5)^T$, for instance.

It is commented in passing that probably the most well-known RR-VGLM is the 'stereotype model' of Anderson (1984). Better called the (rank-1) reduced-rank MLM, it was proposed with the additional constraint of *ordered* elements of $\mathbf{A}$. This is a difficult RR-VGLM estimation problem (e.g., Fernández, Arnold and Pledger, 2016), however it is remarked that plain VGLMs offer a step forward in this direction and a partial solution. For example, one might stipulate that the elements of $\mathbf{A}$ are equally spaced. Another example is when the elements of $\mathbf{A}$ are proportional to the $(M+1)$-quantiles of a standard normal distribution by $a_{j1} \propto \Phi^{-1}(j/(M+1))$ for $j = 1, \ldots, M$. These cases may be handled by setting $\mathbf{H}_k = \mathbf{A}$ for all variables in $\boldsymbol{x}_{[2]}$ and fitting a VGLM with such constraint matrices.

5. Analyses. We analyze the data using VGAM (Yee, 2025) which is available from CRAN. There are often several GAITD regressions, for instance arising from choosing between alteration and inflation, that will yield similar fits. While the analyses follow a list of guidelines and modelling strategies in the Supplementary Materials, some application-specific remarks highlighting certain steps are beneficial at this stage.

(i) We fitted an intercept-only (null) model first. For these, we justify our choices for $\mathcal{S}$ after overlaying the fit onto the data for comparison (Figures 3–6 bar 5). Since we favour choices that are easily and naturally interpreted, we sought a few subgroups that the mixture model helped to identify that had direct bearing on the data and research question.
(ii) To obtain our second model, we regressed a fixed list of covariates (gender, race, marital status, education, and if appropriate, SIA or total passive smoking) and applied AIC-based stepwise regression for variable selection. These produced Tables 2–4. To answer (P1)–(P4) we interpreted the regression coefficients while paying particular attention to their signs. For this, we held other variables fixed when interpreting the effect of one variable. Interpretation for (P1)–(P2) is conditional on subjects being a smoker, and likewise for (P3)–(P4) it is conditional on being a quitter. All analyses are conditional on being less than 80 years old and ex-smokers quitting less than 50 years ago.
(iii) The data were split into a few 50:50 training and test sets and plotted to ascertain the features needing addressing. Since overfitting is effortless, there is a temptation to inflate every spike that protrudes slightly above the main distribution, and likewise deflate every tiny dip. Because we have an aversion to overfitting, our models tend to slightly underfit to avoid numerical problems that can afflict overfitted ones.

(iv) Models having substantial nonparametric inflation were handled by DRR-VGLMs for increased parsimony and interpretability. On occasion, we allowed smoothing by a VGAM, to provide additional insights and exploit the VGLM/VGAM framework versatility.
(v) Since there are four responses, each analysis was necessarily brief. More refined analyses would ideally involve collaboration with a subject matter expert, especially for the identification of subgroups and implications towards public health policy.
(vi) For brevity, parts of the analyses have been placed in the Supplementary Materials.
(vii) Our models were based on studying the spike plots somewhat, therefore all Wald tables such as Table 2 carry a caveat concerning standard inference because of pre-examination. This alludes to a deep and fundamental problem for which there is no easy solution. The AIC was used for goodness of fit. To measure the distance between the GAITD and naive models, we now propose the use of the Kullback–Leibler divergence (KLD; Kullback and Leibler, 1951): $D_{\mathrm{KL}}(f \parallel f_\pi) =$

$$\omega_p \left[\log \frac{\omega_p}{\sum\limits_{u \in \mathcal{A}_p} f_\alpha(u)} + \sum_{a \in \mathcal{A}_p} A_\alpha(a) \, \log \frac{f_\alpha(a)}{f_\pi(a)} \right] + \sum_{s=1}^{|\mathcal{A}_{np}|} \omega_s \, \log \frac{\omega_s}{f_\pi(a)} + \Delta \, (\log \Delta) \, \Pr(y \notin \mathcal{S}) +$$

$$\sum_{i \in \mathcal{I}_p} [\Delta \, f_\pi(i) + \phi_p \, A_\iota(i)] \, \log \left\{ \Delta + \phi_p \, \frac{A_\iota(i)}{f_\pi(i)} \right\} + \sum_{s=1}^{|\mathcal{I}_{np}|} [\Delta \, f_\pi(i_s) + \phi_s] \, \log \left\{ \Delta + \frac{\phi_s}{f_\pi(i_s)} \right\} +$$

$$\sum_{d \in \mathcal{D}_p} [\Delta \, f_\pi(d) - \psi_p \, A_\delta(d)] \, \log \left\{ \Delta - \psi_p \, \frac{A_\delta(d)}{f_\pi(d)} \right\} + \sum_{s=1}^{|\mathcal{D}_{np}|} [\Delta \, f_\pi(d_s) - \psi_s] \, \log \left\{ \Delta - \frac{\psi_s}{f_\pi(d_s)} \right\}$$

as $0 \cdot \log 0 \equiv 0$ by a limit argument, where $\Pr(y \notin \mathcal{S}) = \mathcal{N} - \sum_{t \in \mathcal{T}} f_\pi(t)$, and $A_\alpha(y) = f_\alpha(y) / \sum\limits_{u \in \mathcal{A}_p} f_\alpha(u)$, $A_\iota(y) = f_\iota(y) / \sum\limits_{u \in \mathcal{I}_p} f_\iota(u)$, etc. We used the KLD to measure the total effect of alteration, inflation, truncation and deflation relative to the parent distribution.

5.1. *Smoking initiation age (P1).* It is useful to plot the SIA distributions for the ever-smokers and ex-smokers side-by-side (Figure 3a) for a comparison because there are subtle differences in interpretation between the groups. They are seemingly similar: both are unimodal and right-skewed so that a NBD parent is reasonable, and there is a little heaping at 25, 30, 35 so that $\mathcal{A}_p$ and $\mathcal{I}_p$ are possibilities. A careful examination about age 16 suggests that the probability of quitting for those who started smoking younger than 16 years is less than for those whose smoking onset age is greater than 16. Researchers have conjectured that smokers who begin their habit relatively early end up with a higher tobacco consumption later in life (e.g., Taioli and Wynder, 1991), as well as a higher quitting age if they do manage to quit. This is addressed in Section 5.3. In the following, we show that SIA is well-modelled by a GI mixture distribution comprising three components which can be attributed to a lower-risk subgroup comprising very young and old, a mid-risk young subset, and a high-risk adolescent subgroup.

We chose $\mathcal{I}_p = \{12, 13, 14, 19, 20, 21\}$ and $\mathcal{I}_{np} = \{15, 16, 17, 18\}$ with $f_\pi = f_\iota$. The GI–NB regression is overlaid on the original data in Figure 3(b). The choice of operators and special values describes three subgroups and their justification for these sets is included:

- There is a 'low-risk subgroup' whose age is less than 12 or greater than 21. They can be considered the leftovers from the other subgroups. Its proportion is $\Pr(Y \notin \mathcal{S}) \approx 0.21$.
- f_ι describes a 'mid-risk' age subgroup $\mathcal{I}_p$. It is f_π inflated to describe early and later smokers about the high-risk age subgroup. The proportion $\Pr(Y \in \mathcal{I}_p) \approx 0.32$.

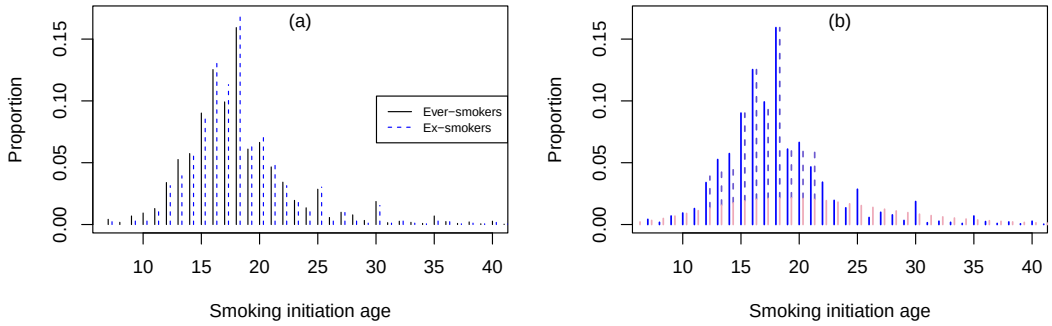

FIG 3. *SIA in the NHANES 2017–2020 data. (a) Ever-smokers and ex-smokers compared side-by-side. (b) Intercept-only GI–NB regression fitted to the ever-smokers. The fitted values have been shifted slightly to the right of the data. It has AIC* $= 19342.7$ *and KLD* $= 0.375$. *Like all spikeplots, the colour scheme is as Fig. 2.*

- The 'high-risk' subgroup has ages $\mathcal{I}_{np}$. These have unstructured probabilities added to f_π and happen to exceed f_ι on either side. The open-ended nature of these probabilities sometimes may be ascribed meaning, and if so it is usually very localized. Here we are modelling the values as they are, viz. at certain ages per se. Making up $\approx 47\%$ of our sample, they correspond to a frenetic teenage smoking uptake: an age group where the largest number of regular smokers start and where there are haphazard effects that cannot be modelled by a NBD. It is not unexpected that 18 has the highest spike because it is the minimum legal age for tobacco purchase. For this reason, the value would be subject to the greatest amount of heaping.
- An alternative is to partition $\mathcal{I}_p$ because those younger than $\mathcal{I}_{np}$ could differ substantially from those older. For this, we tried $\mathcal{I}_p = \{12, 13, 14\}$ and $\mathcal{A}_p = \{19, 20, 21\}$ and this gave some minimal improvement ($\text{AIC} = 19340.4$) but the model is encumbered by using alteration and inflation to explain the spikeplot. We believe it is best to retain inflation as the sole operator because it greatly simplifies the model and interpretation.

The model borrows strength from having a common set of NBD parameters, i.e., $\boldsymbol{\theta}_\pi = \boldsymbol{\theta}_\iota$. This model is quite simple in that $\widehat{\mu}_\pi = \widehat{\mu}_\iota \approx 20.7$ years, i.e., this is the mean age at which smokers start. This can be compared to the sample mean $\overline{y}_{\text{SIA}} \approx 18.4$ years. Also, a naive NB regression is grossly inferior as $\text{AIC} = 20319$.

A practical outcome for public health is to target the high-risk group for prevention education. This might commence slightly earlier, e.g., at ages 13–14.

With covariates and conditional on being a smoker, the model and its coefficients in Table 2 ordered by (7) are

$$\boldsymbol{\eta} \;=\; \big(\log\mu_\pi,\; \log\kappa_\pi,\; \log(\phi_p/\mathcal{N}),\; \log(\phi_{\lceil 15\rceil}/\mathcal{N}),\; \ldots,\; \log(\phi_{\lceil 18\rceil}/\mathcal{N})\big)^T \tag{16}$$

with all but η_1 being intercept-only, and

$$\eta_1 \;=\; \beta^*_{(1)1} + \beta^*_{(1)2}\texttt{race2Others} + \cdots + \beta^*_{(1)9}\texttt{educ} >= \texttt{College graduate}. \tag{17}$$

Table 2 indicates that, conditional on being a smoker, the most highly educated start smoking later relative to those with less education, non-Hispanic whites start smoking earlier, and males take up smoking earlier than females. The model suggests that those widowed/divorced/separated took up smoking later relative to married/partnered persons. There are several possible explanations for this, for example, factors leading to being widowed, divorced or separated such as work stress and poor health could also lead to initiate smoking. If the explanation is at least partially causal then trauma occurring later in life (i.e., these events

are after marriage) predisposing an individual to take up smoking would produce relatively high values of SIA.

The multitude of intercepts in Table 2 demonstrates how easy it is for GAITD regression to produce many coefficients. Furthermore, determining their values on the probability scale requires evaluating the nontrivial softmax function. Nevertheless, some basic information can be gleaned such as the order of intercepts 4–7 implying $\widehat{\phi}_{\lceil 15 \rceil} < \widehat{\phi}_{\lceil 17 \rceil} < \widehat{\phi}_{\lceil 16 \rceil} < \widehat{\phi}_{\lceil 18 \rceil}$.

To summarize our answer to (P1), male non-Hispanic whites aged 15–18 are particularly susceptible to early smoking initiation. Overall, three subgroups were identified and attributed to low, medium and high risk. The latter subgroup might be subjected to the most intensive tobacco control or deterrence policies.

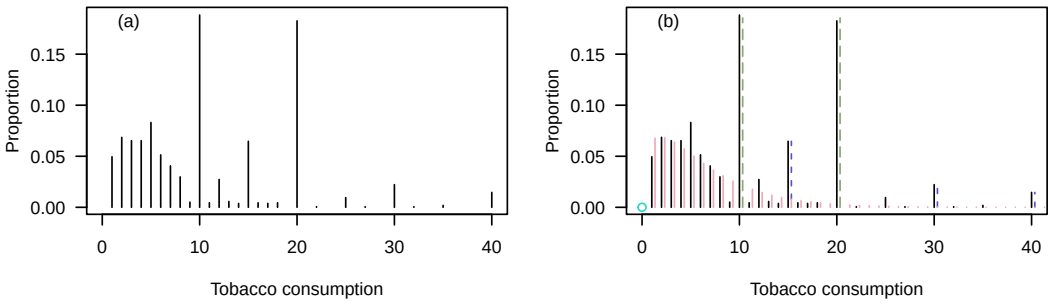

FIG 4. *Daily TC in 1577 current smokers. (a) Raw data; (b) Intercept-only GAIT–NB regression compared to the raw data (RHS shifted and dashed and/or in colour) with AIC* $= 8435.4$ *(cf.* 10521.6 *for naive NB regression) and KLD* $= 1.11$.

5.2. *Tobacco consumption (P2).* The TC distribution (Figure 4a) among current smokers is intriguing and bears a resemblance to Figure 1a. The primary feature is strong heaping in the form of a large and almost equal number of smokers who consume either half or one pack daily against a backdrop of an underlying NB-like distribution accounting for low TC and

TABLE 2

GI–NB regression with covariates for SIA in the NHANES 2017–2020 data. The (stepwise) regression coefficients are ordered by (16) and (17). The AIC $= 19223.9$. *Caveat: SEs, Wald statistics and p-values are approximate only.*

	Estimate	Std. Error	z value	Pr(>|z|)
$\log \mu_\pi$ intercept	2.819	0.043	65.106	0.000
$\log \kappa_\pi$ intercept	2.579	0.065	39.602	0.000
$\log(\phi_p/\mathcal{N})$ intercept	-0.702	0.086	-8.194	0.000
$\log(\phi_{\lceil 15 \rceil}/\mathcal{N})$ intercept	-1.784	0.095	-18.829	0.000
$\log(\phi_{\lceil 16 \rceil}/\mathcal{N})$ intercept	-1.381	0.080	-17.328	0.000
$\log(\phi_{\lceil 17 \rceil}/\mathcal{N})$ intercept	-1.680	0.092	-18.348	0.000
$\log(\phi_{\lceil 18 \rceil}/\mathcal{N})$ intercept	-1.103	0.072	-15.365	0.000
race2Others	0.180	0.020	8.988	0.000
genderMale	-0.057	0.020	-2.906	0.004
maritalNever married	0.019	0.025	0.744	0.457
maritalWidowed/Divorced/Separated	0.082	0.023	3.505	0.000
educ9-11th grade	-0.010	0.044	-0.226	0.821
educHigh school	0.087	0.041	2.116	0.034
educCollege degree	0.127	0.041	3.102	0.002
educ>= College graduate	0.199	0.045	4.376	0.000

having right skew that is punctuated with spikes at other heaped values such as at 0.75, 1.5 and 2 packs. The intercept-only GAIT–NB regression captures the most important features (Figure 4b). For this we chose $\mathcal{T} = \{0\}$ by definition, $\mathcal{A}_{np} = \{10, 20\}$ with a parallelism assumption to make the spikes equal, and $\mathcal{I}_{np} = \{15, 30, 40\}$ to handle the lesser heaping at $\frac{3}{4}, 1\frac{1}{2}$ and 2 packs. Further justification for these selections are as follows. We argue that the main distribution is unimodal and centered around 5. Although it appears to have a long tail, we chose $\mathcal{I}_{np}$ as above rather than $\mathcal{I}_p$ because parametric inflation had a substantial lack of fit. The two big spikes at 10 and 20 appear equal and unrelated to the main distribution. It would be possible to set $\mathcal{D}_{np} = \{9, 11\}$ with a parallelism assumption because they sandwich 10, however the seeping is minor. The inflation is also minor at 5. The scaled parent f_π makes up approximately 52.8% and $\widehat{\mu}_\pi \approx 5.75$ years. In contrast, the overall combo mean is estimated by $\mathrm{E}[Y_{\mathrm{TC}}] \approx 10.94$ which is much higher—it is drawn upwards because of the two large spikes. The sample mean self-reported daily number of cigarettes smoked is about 10.9. In contrast, comparing Figure 4(a) with the global estimate of GBD 2019 Tobacco Collaborators (2021, Fig. 2) shows similarities: a NB-like distribution having a mean of about 10. To gauge the effect of ignoring all GAITD-like features, a naïve NB fitted to TC yields $\widehat{\mu} = 10.89$ and AIC $= 10521.6$ with the latter indicating a very poor fit. The large KLD shows that GAITD-like features are more necessary than with the other three responses.

Before fitting additive models, we performed stepwise regression (effectively backward elimination) because variable selection with VGAMs is heuristic and difficult. The procedure removed marital status and we allowed the covariates to model $\Pr(Y_{\mathrm{TC}} = 10)$ and $\Pr(Y_{\mathrm{TC}} = 20)$ identically. The model and its coefficients in Table 3 ordered by (7) are $\boldsymbol{\eta}^T =$

$$\left(\log \mu_\pi,\ \log \kappa_\pi,\ \log \frac{\omega_{\lceil 10 \rceil}}{\mathcal{N}},\ \log \frac{\omega_{\lceil 20 \rceil}}{\mathcal{N}},\ \log \frac{\phi_{\lceil 15 \rceil}}{\mathcal{N}},\ \log \frac{\phi_{\lceil 30 \rceil}}{\mathcal{N}},\ \log \frac{\phi_{\lceil 40 \rceil}}{\mathcal{N}}\right) \tag{18}$$

with all but η_1, η_3 and η_4 being intercept-only. Hence

$$\eta_1 \ = \ \sum_{k=1}^{6} \beta^*_{(1)k}\, x_k, \qquad \eta_3 \ = \ \eta_4 \ = \ \sum_{k=1}^{6} \beta^*_{(2)k}\, x_k, \tag{19}$$

with `:1` in the computer output denoting coefficients for η_1, etc. The following observations conditional on being a current smoker are made from the table.

- For the covariates, all pairs of significant regression coefficients have the same sign, thus provide nonconflicting evidence.
- The estimated regression coefficients of SIA are negative so that, keeping smoking duration fixed, the data supports the hypothesis that a later SIA is associated with less TC. This affirms the view that prevention education should be aimed at those youngest among the vulnerable, for if they do start smoking later then at least there is a mild trend for them to smoke less initially.
- Keeping SIA fixed, an increased smoking duration is associated with an increased mean TC. This is consistent with the conclusion of the previous point.
- The signs of the remaining estimated regression coefficients are not surprising: (i) non-Hispanic whites appear to smoke more than the other group; (ii) males smoke more than females; (iii) and those exposed to more secondary smoke tend to have a higher TC.

We fitted a VGAM (Figure 5) with

$$\eta_1 = \log\, \mu_\pi \ = \ \beta^*_{(1)1} + f^*_{(1)2}(\mathrm{SIA}) + f^*_{(1)3}(\mathrm{SD}) + \boldsymbol{\beta}_1^T \boldsymbol{x}, \tag{20}$$

$$\eta_j = g(\omega_{\lceil s \rceil}) \ = \ \beta^*_{(j)1} + f^*_{(2)2}(\mathrm{SIA}) + f^*_{(2)3}(\mathrm{SD}) + \boldsymbol{\beta}_j^T \boldsymbol{x}, \tag{21}$$

for $s = 10, 20$ and $j = 3, 4$, where g is the multilogit link applied to model $\Pr(Y_{\text{TC}} = 10)$ and $\Pr(Y_{\text{TC}} = 20)$, and $\eta_3 = \eta_4$. Even though the additive predictors (20)–(21) are quite different quantities, $\widehat{f}^*_{(1)2}$ and $\widehat{f}^*_{(2)2}$ are remarkably similar, as is $\widehat{f}^*_{(1)3}$ to $\widehat{f}^*_{(2)3}$ which is reassuring. Collectively, it indicates that the following conclusions are valid across the entire TC distribution.

- In the first row of plots, the decreasing linear trends with respect to SIA suggests that TC declines with SIA. There is so much uncertainty in the functional form that linear effects were chosen by the automatic smoothing parameter selection algorithm based on O-splines (Wand and Ormerod, 2008).
- The second row of plots show monotonically increasing trends that level off. The greatest TC increase seems immediately after regular smoking begins and becomes almost constant about 30–40 years afterwards.

Together, the VGAM suggests that those taking up smoking later in life, compared to those taking it up earlier in life, smoke less initially, and that daily tobacco consumption tends to increase the longer the dependence persists but with a levelling off around 35 years later. These results concur with Pierce (2022) who advocated reducing smoking intensity (cigarettes per day) as one of the goals for tobacco control programs, especially for those younger than 35 years.

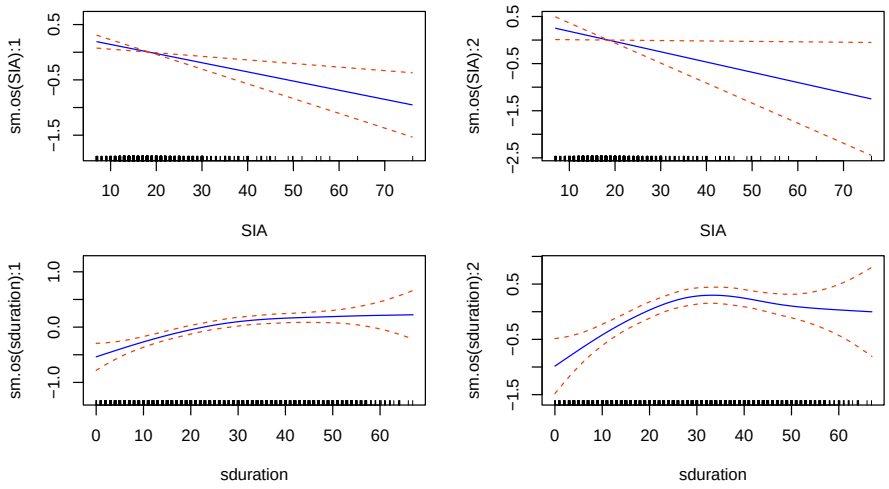

FIG 5. *Estimated component functions (6) of a GAIT–NB VGAM fitted to tobacco consumption in current smokers. The dashed curves are ±2 pointwise SE bands. The covariates are smoking initiation age and smoking duration. The LHS and RHS plots are (20)–(21) respectively: from top-left going clockwise, they are* $\widehat{f}^*_{(1)2}(\text{SIA})$, $\widehat{f}^*_{(2)2}(\text{SIA})$, $\widehat{f}^*_{(2)3}(\text{SD})$, $\widehat{f}^*_{(1)3}(\text{SD})$.

5.3. *Smoking duration (P3).* Being the difference between a heaped and an unheaped variable, the smoking duration distribution among ex-smokers appears ragged (Figure 6a). After a naïve NB was fit, it was evident that any conventional parametric distribution would struggle and that the shape was humped for the first few values. We remark that these excess individuals belong to a group of *quick quitters*. Hence we fitted an intercept-only GI–NB regression with a parallelism assumption applied to $\mathcal{I}_{np} = \{0, 1, \ldots, 5\}$ to simplify the spikes in Figure 6(b) to have equal height. This resulted in the AIC decreasing by 44.9. Contrary to

TABLE 3

GAIT–NB regression for current tobacco consumption after stepwise regression. The order of the intercepts is given in (18) and (19). The AIC $= 8216.8$. *Caveat: SEs, Wald statistics and p-values are approximate only.*

	Estimate	Std. Error	z value	Pr(>|z|)
$\log \mu_\pi$ intercept	1.773	0.137	12.923	0.000
$\log \kappa_\pi$ intercept	0.795	0.088	9.002	0.000
$\log(\omega_{\lceil 10,20 \rceil}/\mathcal{N})$ intercept	−0.651	0.262	−2.482	0.013
$\log(\phi_{\lceil 15 \rceil}/\mathcal{N})$ intercept	−2.305	0.123	−18.778	0.000
$\log(\phi_{\lceil 30 \rceil}/\mathcal{N})$ intercept	−3.281	0.182	−17.995	0.000
$\log(\phi_{\lceil 40 \rceil}/\mathcal{N})$ intercept	−3.636	0.215	−16.944	0.000
SIA:1	−0.018	0.005	−3.645	0.000
SIA:2	−0.024	0.010	−2.369	0.018
sduration:1	0.012	0.002	6.225	0.000
sduration:2	0.013	0.004	3.497	0.000
race2Others:1	−0.408	0.061	−6.638	0.000
race2Others:2	−0.751	0.111	−6.784	0.000
genderMale:1	0.203	0.059	3.404	0.001
genderMale:2	−0.011	0.110	−0.103	0.918
totpassive:1	0.148	0.024	6.130	0.000
totpassive:2	0.085	0.046	1.852	0.064

this, allowing each spike to be unconstrained would result in a overfitted model whose values fit perfectly. A KLD of 0.065 indicates the modification on the NBD is moderate.

Table 4 summarizes the covariate model after stepwise regression removed gender and race. The model has

$$\boldsymbol{\eta} = \left(\log \mu_\pi,\ \log \kappa_\pi,\ \log\left(\phi_{\lceil 0 \rceil}/\mathcal{N}\right),\ \ldots,\ \log\left(\phi_{\lceil 5 \rceil}/\mathcal{N}\right)\right)^T \tag{22}$$

with only η_2 being intercept-only and $\eta_3 = \eta_4 = \cdots = \eta_8$ to model all quick quitters symmetrically. The following findings, conditional on being an ex-smoker and fixing the other variables, are made.

- A shorter mean SD is associated with an increasing SIA because of the negative coefficient in η_1 and positive coefficient $\widehat{\beta}^*_{(2)2}$. The latter means that the excess chances of quitting within 5 years increases with SIA. These help answer (P3), viz. that those who initiate smoking later in life have a higher probability of becoming quick quitters in the sense that the number of years spent smoking is on average less than those who started smoking at a younger age.

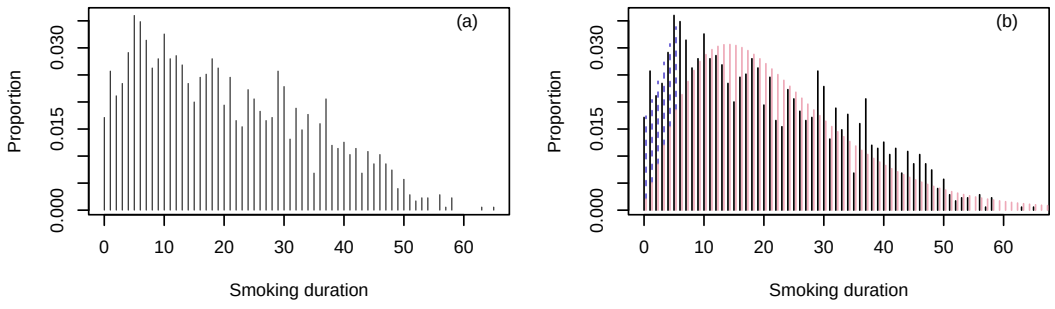

FIG 6. *Smoking duration of ex-smokers in the NHANES 2017–2020 data. The value at 0 are participants who started smoking within 6 months of data collection. (a) Raw data; (b) Intercept-only GI–NB regression (AIC* $=$ 13827.9*) and compared to (a). Quick quitters are the first six values having inflation (dashed lines of equal height).*

- While individuals who are widowed/divorced/separated tend to smoke longer on average than those married/partnered, the never married tend to quit the fastest. A possible explanation for the latter is that individuals in dual-smoker couples typically report low motivation to quit smoking (vanDellen et al., 2019).
- There is a tendency for SD to decrease with increasing education.

Although these findings are as anticipated, monotonicity of the $\widehat{\beta}^{*}_{(j)k}$ is not observed for education in η_1 and η_3. We now fit a more interpretable DRR-VGLM where the monotonicity in education is stronger than in the case of the VGLM fit.

5.3.1. *DRR-VGLM analysis.* Fitting (15) with $\boldsymbol{x}_{[1]} = 1$, ν is defined by a linear combination of $(\texttt{SIA}, \texttt{marital}, \texttt{educ})^T$, with marital status and education expanding out to several indicator variables. Then the latent variable is used to model $\log \mu_\pi$ and the $\phi_{\lceil s \rceil}$ $(s = 0, 1, \ldots, 5)$ simultaneously. Our reduced rank regression has structure as the values of s are treated equally: the constraint matrices for the intercept and $\mathbf{A}$ (i.e., (5) and (14)) are

$$\mathbf{H}_1 = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{O}_{6\times 2} & \mathbf{1}_6 \end{pmatrix}, \quad \mathbf{H}_{\text{A1}} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ \mathbf{0}_6 & \mathbf{1}_6 \end{pmatrix}.$$

Ordinarily, RR-VGLMs would have an estimate for each element of $\mathbf{A}$ and therefore be prone to overfitting.

From Table 5 we have $a_{11} \equiv 1$, $\widehat{a}_{31} = \widehat{a}_{41} = \cdots = \widehat{a}_{81} \approx -2.294$, $\log \widehat{\kappa}_\pi \approx 1.188$, etc. In contrast with Table 4, the single effect of education is almost strictly monotonic, suggesting a more parsimonious model. Each estimated coefficient of $\mathbf{C}$ is very interpretable: ν is decreasing with increasing SIA and increasing education, and has a positive loading for those who are widowed/divorced/separated. Thus holistically, a quitter having a higher value of ν is more 'unhealthy' because smoking commences at a lower age, s/he has lower education and is widowed/divorced/separated. That a_{11} and $\widehat{a}_{31}$ have opposite signs implies that mean smoking duration and the excess probability of being a quick quitter move in opposite directions as the latent variable changes. This makes sense, e.g., with decreasing ν there is a

TABLE 4

GI–NB regression with covariates after stepwise regression for smoking duration in the NHANES 2017–2020 data. The elements of $\boldsymbol{\eta}$ are enumerated by (22). Caveat: SEs, Wald statistics and p-values are approximate only.

	Estimate	Std. Error	z value	Pr(>|z|)
$\log \mu_\pi$ intercept	3.674	0.083	44.458	0.000
$\log \kappa_\pi$ intercept	1.199	0.056	21.541	0.000
$\log(\phi_{\lceil 0,\ldots,5 \rceil}/\mathcal{N})$ intercept	−5.550	0.547	−10.142	0.000
SIA:1	−0.029	0.004	−8.037	0.000
SIA:2	0.071	0.018	4.008	0.000
maritalNever married:1	−0.060	0.050	−1.190	0.234
maritalNever married:2	0.560	0.245	2.282	0.022
maritalWidowed/Divorced/Separated:1	0.197	0.037	5.400	0.000
maritalWidowed/Divorced/Separated:2	−1.070	0.346	−3.091	0.002
educ9-11th grade:1	0.010	0.071	0.138	0.890
educ9-11th grade:2	0.082	0.539	0.153	0.879
educHigh school:1	−0.026	0.063	−0.406	0.685
educHigh school:2	0.365	0.467	0.781	0.435
educCollege degree:1	−0.110	0.061	−1.807	0.071
educCollege degree:2	0.490	0.454	1.079	0.281
educ>= College graduate:1	−0.347	0.067	−5.215	0.000
educ>= College graduate:2	0.041	0.522	0.079	0.937

protective effect due to mean smoking duration decreasing coupled with the quick quitting probability increasing.

Compared to a RR-VGLM where $\mathbf{A} = (1, 0, a_{31}, \ldots, a_{81})^T$ and AIC = 13629.2, the DRR-VGLM has an AIC that is 5.0 less. Thus there is real benefit gained from the DRR-VGLM in terms of interpretative ease and goodness of fit.

TABLE 5

DRR-VGLM GI–NB regression in ex-smokers, for smoking duration, in the NHANES 2017–2020 data. The rows have been partially partitioned into estimates for $\mathbf{A}$, $\mathbf{B}_1$ *and* $\mathbf{C}$ *respectively, so the first row is for* $\widehat{a}_{31}$–$\widehat{a}_{81}$. *The AIC is 13624.1. Caveat: SEs, Wald statistics and p-values are approximate only.*

	Estimate	Std. Error	z value	Pr(>|z|)
$a_{31}, \ldots, a_{81}$	−2.294	0.661	−3.467	0.000
$\log \mu_\pi$ intercept	3.680	0.084	43.726	0.000
$\log \kappa_\pi$ intercept	1.188	0.056	21.206	0.000
$\log(\phi_{[0,\ldots,5]}/\mathcal{N})$ intercept	−5.483	0.383	−14.308	0.000
SIA	−0.029	0.004	−7.938	0.000
maritalNever married	−0.103	0.042	−2.468	0.007
maritalWidowed/Divorced/Separated	0.225	0.034	6.688	0.000
educ9-11th grade	0.004	0.064	0.055	0.478
educHigh school	−0.046	0.057	−0.811	0.209
educCollege degree	−0.127	0.055	−2.306	0.011
educ>= College graduate	−0.291	0.064	−4.565	0.000

5.3.2. *Fitting a VGAM.* Because `SIA` is of central interest and statistically significant, we fit a simplified additive model with this covariate only. The VGAM (cf. (22) and (6)) is

$$\eta_1 = \beta^*_{(1)1} + f^*_{(1)2}(\texttt{SIA}), \qquad \eta_2 = \beta^*_{(2)1}, \qquad \eta_j = \beta^*_{(3)1}, \quad j = 3, \ldots, 8, \tag{23}$$

and plotted in Figure 7. The fitted component function (6) on the LHS includes pointwise ±2SE bands to provide variability information. Changes in this function may be 'magnified' by considering its first derivative shown on the right. For instance, determining the SIA where SD changes the least is more easily seen from the right panel. We comment that:

- Plot (a) shows an almost linear decreasing trend suggesting that the earlier one starts smoking, the longer the period spent smoking. However, because it seems most rapidly decreasing for SIA $\leq$ 20 years (plot (b)), this suggests that one should redouble efforts to stop smoking starting before age 20.
- The SE-bands are wide at the boundaries and indicate that there is considerable uncertainty at extreme SIA values, hence a need to refrain from over-interpreting there. For instance, the decrease on the RHS of plot (b) is almost surely spurious and this is further supported by the rugplot showing a lack of data in that region. Nevertheless, the public health implication is obvious: those taking up smoking while very young are the most vulnerable, therefore they should be targetted for smoking prevention education.

Together, the smooth suggests that those who start smoking at very young ages tend to smoke much longer. To answer (P3), (23) and Figure 7 indicate that the lowest SIA is associated with highest SD, hence there is a strong case for smoking prevention education aimed at the early- or pre-teenage years.

5.4. *Smoking cessation age (P4).* The SCA distribution among ex-smokers, which is similar to smoking duration due to $Y_{\rm SCA} = Y_{\rm SIA} + Y_{\rm SD}$, is unimodal (Figure 8a) and viewed as a NB distribution augmented with a hump on its top LHS. As mentioned in Section 1.4, heaping and seeping in Figure 8(a) is rather subdued because SCA was calculated by subtracting the self-reported quitting duration from the participant's age, hence any heaping is diluted, e.g., the small repeating pattern of seepage in the 30s, 40s and 60s age ranges.

The hump covers the range $18 \le {\rm SCA} \le 32$ approximately and these individuals are assigned the name '*young quitters*'. There is a seemingly large spike at 25, however, plots of the training/test data revealed that this spike could frequently disappear so that treating it differently from its neighbors would be overfitting. Like quick quitters, young quitters are of interest because they might be an appropriate audience of a public health initiative.

We chose a GIT–NB model with $\mathcal{T} = \{0, \ldots, 6\}$ because the minimum SIA was 7 and truncating beyond 80 years is negligible. Also, $\mathcal{I}_{np} = \{18, \ldots, 32\}$ with a parallelism assumption for $g(\phi_s)$ to create a band of spikes having a constant height. A plausible explanation is that because SIA is peaked around 15–20, there are many young ex-smokers having a short smoking duration. We conclude from Figure 8(b) that to a large extent, SCA is negative binomial distributed with an even band of probabilities added on at ages 18 to 32 to account for the young quitters. That is, while smoking initiation is largely in the teenage–young adult years, quitting is lagged and peaks at around 20–30 years of age. The band height corresponds to a probability of 1.5%, and when summed this is a substantial 22.6%. An abrupt drop-off after 32 years is noticeable and its reason is unknown. The scaled parent f_π makes up approximately 60.7% and its mean is $\widehat{\mu}_\pi \approx 43$ years. In contrast, the overall combo mean is estimated by ${\rm E}(Y_{\rm SCA}) \approx 38.7$ years—less because of the hump. The effect of ignoring all GAITD-like features is an inferior naïve NB with $\widehat{\mu} = 38.6$ and AIC $= 13968.4$.

The public health implication is that there is good reason for intensifying cessation campaigns for those less than age 40 or so because most quitting occurs before that age and the potential benefits are greatest then (Pierce, 2022). Indeed, the median of the GAIT–NB model is only slightly higher than 40 years. These results are also in line with Thomson et al. (2022) who concluded that quitting smoking, particularly at younger ages, was associated with substantial reductions in the relative excess mortality associated with continued smoking.

With covariates, stepwise regression chose marital status and education for inclusion (AIC $= 13735.8$). From this, young quitters tend to be single and definitely not widowed/divorced/separated. However, most of the constituent p-values of `educ` were nonsignificant. This problem was overcome by the DRR-VGLM.

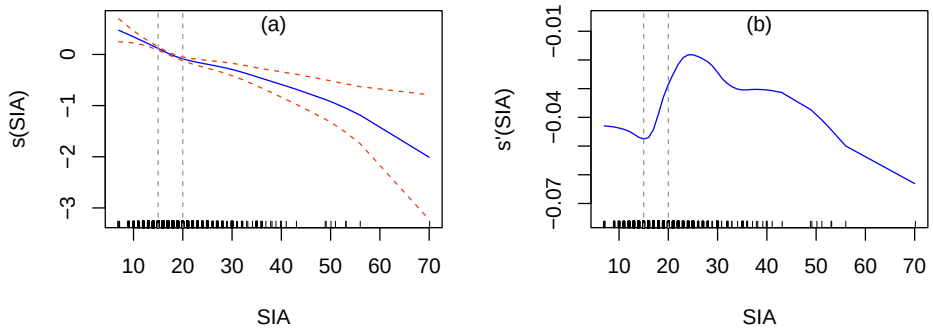

FIG 7. *(a) Estimated component function* $\widehat{f}^{*}_{(1)2}({\rm SIA})$ *from a GI–NB VGAM (23) fitted to smoking duration in ex-smokers. (b) Its first derivative. Vertical dashed lines at ages* 15 *and 20 years have been placed for reference. From (23), the plots concern* $\eta_1 = \log \mu_\pi$.

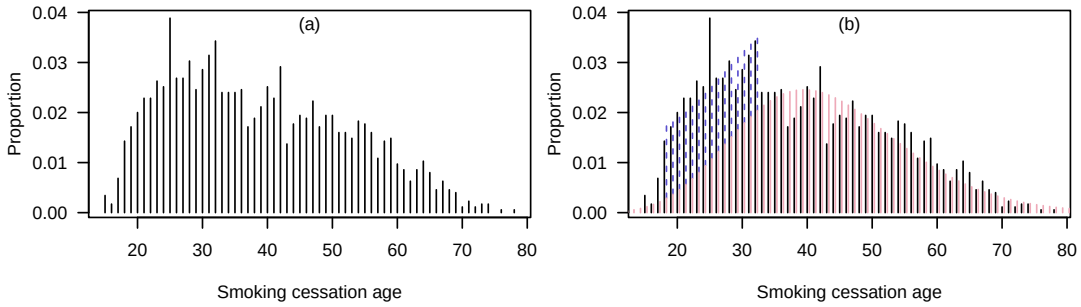

FIG 8. *SCA of* 1750 *quitters from NHANES 2017–2020. (a) Raw data; (b) Intercept-only GIT–NB regression overlaid on (a) and having AIC* $= 13856.6$ *and KLD* $= 0.14$.

TABLE 6
DRR-VGLM GIT–NB regression in ex-smokers, for SCA, in the NHANES 2017–2020 data. The AIC $= 13736.1$. *Caveat: SEs, Wald statistics and p-values are approximate only*

	Estimate	Std. Error	z value	Pr(>|z|)
$a_{3,1}, \ldots, a_{17,1}$	−5.548	1.960	−2.830	0.002
$\log \mu_\pi$ intercept	3.773	0.025	148.242	0.000
$\log \kappa_\pi$ intercept	2.768	0.070	39.682	0.000
$\log(\phi_{\lceil 0,\ldots,5 \rceil}/\mathcal{N})$ intercept	−4.151	0.179	−23.213	0.000
maritalNever married	−0.043	0.020	−2.200	0.014
maritalWidowed/Divorced/Separated	0.116	0.019	6.070	0.000
educ9-11th grade	0.001	0.030	0.041	0.484
educHigh school	−0.018	0.027	−0.660	0.255
educCollege degree	−0.058	0.026	−2.192	0.014
educ>= College graduate	−0.139	0.033	−4.157	0.000

5.4.1. *DRR-VGLM analysis.* Table 6 admits the following interpretation.

- With similarities with SD, $\widehat{a}_{3,1} = \cdots = \widehat{a}_{17,1} \approx -5.548 < 0$ shows that the mean SCA and inflation probability of being a young quitter move in opposite directions.
- As $\widehat{\nu}$ decreases with increasing education, and has a positive loading for those who are widowed/divorced/separated, higher values of the latent variable are more 'unhealthy'.
- With increasing education, the trend is monotonic and the p-values decrease, so there is greater ability to detect differences between less versus more education levels.
- With the married as baseline, both widowed/divorced/separated and singles differ significantly and have opposite signs. This is more conclusive than the VGLM.

All these are of no surprise at all. The AIC of the DRR-VGLM only differs from the stepwise VGLM by 0.3 and is much more interpretable.

As a final diagnostic, we simulated two data sets from the fitted model in Fig. 9. The plot shows the DRR-VGLM adapts to the overall distribution well. The bottom plot shows how ordinary sampling variation can produce a large spike at random (here at 31). Thus it confirms that treating 25 the same as its neighbours is not unreasonable.

6. Discussion. Globally over the past three decades it has been estimated that more than 200 million deaths have been caused by smoking tobacco use (GBD 2019 Tobacco Collaborators, 2021). In the US, smoking remains the leading preventable cause of death (Smoking Cessation (OSG), 2020) as reflected by half a million deaths annually (Thomson et al., 2022). Partly because of these, smoking data is perhaps the most commonly collected data type in

addiction science. This paper has addressed issues surrounding heaped and seeped data in the context of smoking studies and utilized a newly developed regression method for investigating three generic problems fundamental to any simple addiction study with count responses. Heaped and seeped data will always remain a formidable and challenging problem because it can arise in so many shapes and sizes, however we opine that GAITD regression is sufficiently flexible to make inroads in this problem area in the hands of competent practitioners.

As seen for (P3)–(P4), when there are covariates a GAITD regression model can benefit from reduced rank regression. DRR-VGLMs boost the benefits of this dimension reduction with structure but do require careful attention when defining the required constraint matrices. Often the latent variables are readily interpretable and this was seen in Sections 5.3.1 and 5.4.1.

There is room for many refinements not illustrated here. For example, some individuals might be described as recreational smokers whose daily TC for weekdays and weekends differ significantly e.g., `smoking` in openintro (Çetinkaya Rundel et al., 2022). Another example stems from Section 5.4 where our analyses were conditional on successful quitting. In practice the relationship between SIA and (final) SCA is more complex as it depends on the probability of quitting and there is the possibility of multiple cycles of quitting and relapse, e.g., about 30–50% of US smokers attempt to quit in any given year and success rates are only about 7.5% (Creamer et al., 2019). A survival analysis approach could be complementary for teasing out the SIA–SCA relationship. Having similarities with a multivariate GLM with mixture model characteristics, we demonstrated the straightforwardness of GAITD regression by solving basic problems adequately in spite of spikes, dips, bimodality and truncation. For GAITD regression itself, one refinement needed is to handle incorrect y values such as false 0s (Lewbel, 2000). They occur commonly in ecology and some comments are given in Yee and Ma (2024).

More generally, one major and overarching purpose of GAITD regression is to allow inference on the parent distribution by adjusting for or removing unwanted values by alteration, shaving off the excesses or spikes by inflation, filling in deficiencies or dips by deflation, and acknowledging impossible values by truncation. Sometimes the altered, inflated and deflated mixture distributions may be of interest in their own right, though many users will want to constrain them to equal the parent to borrow strength and to avoid numerical problems. Also, the multinomial logit model probabilities may be of interest too.

This work is ongoing, with possible extensions including the development of GAITD regression to continuous distributions. As SIA and SCA are time measurements, they are

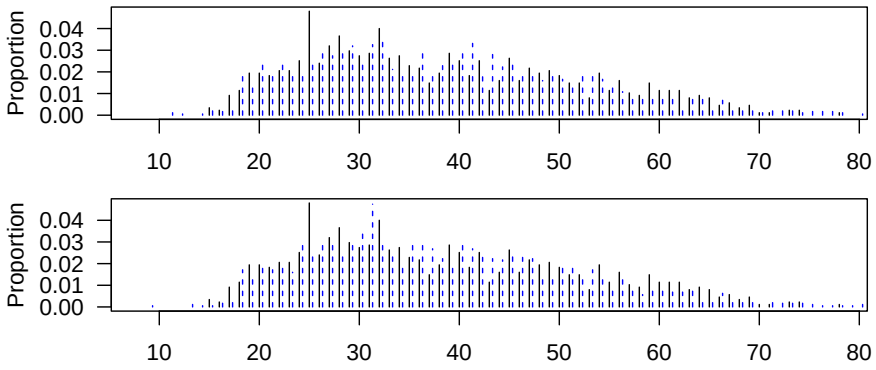

FIG 9. *Smoking cessation age in ex-smokers: simulated data (shifted slightly to the RHS and dashed blue) from the fitted DRR-VGLM, placed adjacent to the final test set (LHS solid black lines).*

strictly continuous so there is a need to develop *concrete* GAITD regression that shares both *con*tinuous and dis*crete* properties for handling heaping and seeping at discrete values from an otherwise continuous distribution. Other future work includes applying the method to other heaped NHANES addiction variables which are easily identified by spikeplotting, e.g., consumption of alcohol and marijuana, number of sexual partners (Pruim, 2015). Alcohol studies would be our first choice, as this is perhaps the next most common data type in large-scale addiction studies whose interventions have been integrated with smoking cessation (Villanti et al., 2020).

Acknowledgements. We thank the six reviewers for many invaluable comments, and Liza Bolton and Rolf Turner for very helpful proof-reading.

SUPPLEMENTARY MATERIAL

Appendix

Contains '*Modelling Strategies and Guidelines for GAITD Regression*', a more detailed analysis of the data, and a simulation study.

REFERENCES

ALI, F. R. M., AGAKU, I. T., SHARAPOVA, S. R., REIMELS, E. A. and HOMA, D. M. (2020). Onset of regular smoking before age 21 and subsequent nicotine dependence and cessation behavior among US adult smokers. *Preventing Chronic Disease* **17** 1–6.

ALLEN, C. M., GRIFFITH, S. D., SHIFFMAN, S. and HEITJAN, D. F. (2017). Proximity and gravity: modeling heaped self-reports. *Statist. Med.* **36** 3200–3215.

ANDERSON, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statistics* **22** 327–351.

ANDERSON, J. A. (1984). Regression and ordered categorical variables. *J. Roy. Statist. Soc. Ser. B* **46** 1–30. With discussion.

ARTINGER, F. M., GIGERENZER, G. and JACOBS, P. (2022). Satisficing: Integrating two traditions. *J. Econ. Lit.* **60** 598–635.

BABOR, T. F. (2000). Past as prologue: The future of addiction studies. *Addiction* **95** 7–10.

BAR, H. Y. and LILLARD, D. R. (2012). Accounting for heaping in retrospectively reported event data–a mixture-model approach. *Statist. Med.* **31** 3347–3365.

BROWN, R. A., BURGESS, E. S., SALES, S. D., WHITELEY, J. A., EVANS, D. M. and MILLER, I. W. (1998). Reliability and validity of a smoking timeline follow-back interview. *Psych. Addict. Behav.* **12** 101.

BUCKLAND, P. R. (2008). Will we ever find the genes for addiction? *Addiction* **103** 1768–1776.

BURA, E., DUARTE, S., FORZANI, L., SMUCLER, E. and SUED, M. (2018). Asymptotic theory for maximum likelihood estimates in reduced-rank multivariate generalized linear models. *Statistics* **52** 1005–1024.

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Second ed. Chapman & Hall/CRC, Boca Raton, FL, USA.

ÇETINKAYA RUNDEL, M., DIEZ, D., BRAY, A., KIM, A. Y., BAUMER, B., ISMAY, C., PATERNO, N. and BARR, C. (2022). openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs. R package version 2.4.0.

CHANG, W., CHENG, J., ALLAIRE, J., SIEVERT, C., SCHLOERKE, B., XIE, Y., ALLEN, J., MCPHERSON, J., DIPERT, A. and BORGES, B. (2023). shiny: Web Application Framework for R. R package version 1.8.0.

CHOI, S. H. (2022). A systematic review and narrative summary of couple-based smoking cessation interventions. *J. Soc. Person. Relat.* **39** 1901–1916.

CHRISTELIS, D. and SANZ-DE GALDEANO, A. (2009). Smoking Persistence Across Countries: An Analysis Using Semi-Parametric Dynamic Panel Data Models with Selectivity Technical Report, Forschungsinstitut zur Zukunft der Arbeit (Institute for the Study of Labor), Bonn, Germany.

GBD 2019 TOBACCO COLLABORATORS (2021). Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019. *The Lancet* **397** 2337–2360.

COX, N. J. (2004). Speaking Stata: Graphing distributions. *Stata J.* **4** 66–88.

CRAWFORD, F. W., WEISS, R. E. and SUCHARD, M. A. (2015). Sex, lies and self-reported counts: Bayesian mixture models for heaping in longitudinal count data via birth-death processes. *Ann. Appl. Stat.* **9** 572–596.

CREAMER, M. R., WANG, T. W., BABB, S., CULLEN, K. A., DAY, H., WILLIS, G., JAMAL, A. and NEFF, L. (2019). Tobacco product use and cessation indicators among adults—United States. *Morbidity and Mortality Weekly Report* **68** 1013–1019.

CUMMINGS, T. H., HARDIN, J. W., MCLAIN, A. C., HUSSEY, J. R., BENNETT, K. J. and WINGOOD, G. M. (2015). Modeling heaped count data. *Stata J.* **15** 457–479.

DOLL, R. and HILL, A. B. (1950). Smoking and carcinoma of the lung; preliminary report. *Brit. Med. J.* **2** 739–748.

DSM-5-TR (2022). *Diagnostic and Statistical Manual of Mental Disorders*, 5th, text revision ed. American Psychiatric Association, Arlington, VA, USA.

EUM, Y. H., KIM, H. J., BAK, S., LEE, S.-H., KIM, J., PARK, S. H., HWANG, S. E. and OH, B. (2022). Factors related to the success of smoking cessation: A retrospective cohort study in Korea. *Tob. Induc. Dis.* **20** 1–10.

FERNÁNDEZ, D., ARNOLD, R. and PLEDGER, S. (2016). Mixture-based clustering for the ordered-stereotype model. *Computational Statistics & Data Analysis* **93** 46–75.

FESSLER, J. A. (1991). Nonparametric fixed-interval smoothing with vector splines. *IEEE Transactions on Signal Processing* **39** 852–859.

FORSTER, M. and JONES, A. M. (2001). The role of tobacco taxes in starting and quitting smoking: Duration analysis of British data. *J. Roy. Statist. Soc. Ser. A* **164** 517–547.

GOODMAN, A. (1990). Addiction: definition and implications. *Brit. J. Addict.* **85** 1403–1408.

HARRIS, M. N. and ZHAO, X. (2007). A zero-inflated ordered probit model, with an application to modelling tobacco consumption. *J. Economet.* **141** 1073–1099.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

HILBE, J. M. (2011). *Negative Binomial Regression*, Second ed. Cambridge University Press, Cambridge, UK; New York, USA.

HU, C. Y., IKSANOV, A. M., LIN, G. D. and ZAKUSYLO, O. K. (2006). The Hurwitz zeta distribution. *Australian & New Zealand Journal of Statistics* **48** 1–6.

ICD-11 (2019). *International Statistical Classification of Diseases and Related Health Problems*, 11th ed. World Health Organization, Geneva, Switzerland.

IZENMAN, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis* **5** 248–264.

JONES, A. M. (1989). A double-hurdle model of cigarette consumption. *J. Appl. Economet.* **4** 23–39.

JORGENSEN, B. N., PATRICK, P. H. and SODERSTROM, N. S. (2020). Heaping of Executive Compensation. *J. Managem. Acc. Res.* **32** 177–201.

JUNG, H. Y., CHOI, H. and PARK, T. (2018). Fuzzy heaping mechanism for heaped count data with imprecision. *Soft Comput.* **22** 4585–4594.

KHUDER, S. A., DAYAL, H. D. and MUTGI, A. B. (1999). Age at smoking onset and its effect on smoking cessation. *Addictive Behaviors* **24** 673–677.

KLESGES, R. C., DEBON, M. and RAY, J. W. (1995). Are self-reports of smoking rate biased? Evidence from the Second National Health and Nutrition Examination Survey. *J. Clin. Epidem.* **48** 1225–1233.

KRINSKY, R. and NELSON, T. O. (1985). The feeling of knowing for different types of retrieval failure. *Acta Psych.* **58** 141–158.

KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22** 79–86.

LAWLESS, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* **15** 209–225.

LEWBEL, A. (2000). Identification of the binary choice model with misclassification. *Econom. Theory* **16** 603–609.

LEWIS-ESQUERRE, J. M., COLBY, S. M., O'LEARY TEVYAW, T., EATON, C. A., KAHLER, C. W. and MONTI, P. M. (2005). Validation of the timeline follow-back in the assessment of adolescent smoking. *Drug and Alcohol Dependence* **79** 33–43.

MACMAHON, S., NORTON, R., JACKSON, R., MACKIE, M. J., CHENG, A., VANDER HOORN, S., MILNE, A. and MCCULLOCH, A. (1995). Fletcher Challenge-University of Auckland Heart & Health Study: Design and Baseline Findings. *N. Z. Med. J.* **108** 499–502.

MAHMUD, M. S., FANG, H., CARREIRO, S., WANG, H. and BOYER, E. W. (2019). Wearables technology for drug abuse detection: A survey of recent advancement. *Smart Health* **13** 100062.

MAK, K. K., LEE, K. and PARK, C. (2019). Applications of machine learning in addiction studies: A systematic review. *Psych. Res.* **275** 53–60.

SMOKING CESSATION (OSG) (2020). Smoking Cessation (Office of the Surgeon General), Centers for Disease Control and Prevention, Atlanta, GA, USA.

PIERCE, J. D. (2022). Quitting smoking by age 35 years—A goal for reducing mortality. *JAMA Network Open* **5** e2231487.

POWERS, S., HASTIE, T. and TIBSHIRANI, R. (2018). Nuclear penalized multinomial regression with an application to predicting at bat outcomes in baseball. *Statistical Modelling* **18** 1–23.

PRUIM, R. (2015). NHANES: Data from the US National Health and Nutrition Examination Study. R package version 2.1.0.

QIU, D., CHEN, T., LIU, T. and SONG, F. (2020). Smoking cessation and related factors in middle-aged and older Chinese adults: Evidence from a longitudinal study. *PLoS ONE* **15** e0240806.

REINSEL, G. C., VELU, R. P. and CHEN, K. (2022). *Multivariate Reduced-Rank Regression: Theory and Applications*, Second ed. Springer-Verlag, New York, USA.

RIBISL, K. M. and MILLS, S. D. (2019). Explaining the rapid adoption of Tobacco 21 policies in the United States. *Am. J. Public Health* **109** 1483–1485.

RICHARDS, F. S. G. (1961). A method of maximum-likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **23** 469–475.

SUSSMAN, S. and SUSSMAN, A. N. (2011). Considering the definition of addiction. *Inter. J. Environ. Res. Pub. Health* **8** 4025–4038.

TAIOLI, E. and WYNDER, E. L. (1991). Effect of the age at which smoking begins on frequency of smoking in adulthood. *New Eng. J. Med.* **325** 968–969.

THOMSON, B., EMBERSON, J., LACEY, B., LEWINGTON, S., PETO, R., JEMAL, A. and ISLAMI, F. (2022). Association between smoking, smoking cessation, and mortality by race, ethnicity, and sex among US adults. *JAMA Network Open* **5** e2231480.

VANDELLEN, M. R., LEWIS, M. A., TOLL, B. A. and LIPKUS, I. M. (2019). Do couple-focused cessation messages increase motivation to quit among dual-smoker couples? *Journal of smoking cessation* **14** 95–103.

VILLANTI, A. C., WEST, J. C., KLEMPERER, E. M., GRAHAM, A. L., MAYS, D., MERMELSTEIN, R. J. and HIGGINS, S. T. (2020). Smoking-cessation interventions for U.S. young adults: Updated systematic review. *Amer. J. Prev. Med.* **59** 123–136.

WAND, M. P. and ORMEROD, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics* **50** 179–198.

WANG, H. and HEITJAN, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Statist. Med.* **27** 3789–3804.

WANG, H., SHIFFMAN, S., GRIFFITH, S. D. and HEITJAN, D. F. (2012). Truth and memory: Linking instantaneous and retrospective self-reported cigarette consumption. *Ann. Appl. Stat.* **6** 1689–1706.

WEST, R. (2016). Using Bayesian analysis for hypothesis testing in addiction science. *Addiction* **111** 3–4.

WILSON, N., HOEK, J., NGHIEM, N., SUMMERS, J., GROUT, L. and EDWARDS, R. (2022). Modelling the impacts of tobacco denicotinisation on achieving the Smokefree 2025 goal in Aotearoa New Zealand. *N. Z. Med. J.* **135** 65–76.

WOLFF, J. and AUGUSTIN, T. (2003). Heaping and its consequences for duration analysis: A simulation study. *AStA Adv. Statist. Anal.* **87** 59–86.

WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R*, Second ed. Chapman & Hall/CRC, London.

YEE, T. W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer, New York, USA.

YEE, T. W. (2020). The VGAM package for negative binomial regression. *Aust. N. Z. J. Stat.* **62** 116–131.

YEE, T. W. (2025). VGAM: Vector Generalized Linear and Additive Models R package version 1.1-13.

YEE, T. W. and FRIGAU, L. (2025). Heaped data: A review and a solution. *In preparation*.

YEE, T. W. and GRAY, J. (2025). VGAMdata: Data Supporting the 'VGAM' Package. R package version 1.1-13.

YEE, T. W. and HASTIE, T. J. (2003). Reduced-rank Vector Generalized Linear Models. *Statist. Modelling* **3** 15–41.

YEE, T. W. and MA, C. (2024). Generally altered, inflated, truncated and deflated regression. *Statist. Sci.* **39** 568–588.

YEE, T. W. and WILD, C. J. (1996). Vector Generalized Additive Models. *J. Roy. Statist. Soc. Ser. B* **58** 481–493.

ZIPF, G., CHIAPPA, M., PORTER, K. S. and ET AL. (2013). National Health and Nutrition Examination Survey: Plan and Operations 1999–2010. *National Center for Health Statistics. Vital and Health Statistics* **1**.

ZOU, Z., WANG, H., D'OLEIRE UQUILLAS, F., WANG, X., DING, J. and CHEN, H. (2017). *Definition of Substance and Non-substance Addiction* In *Substance and Non-substance Addiction* 21–41. Springer Singapore, Singapore.