2015

Latent Variable Approach to Elicit Continuous Toxicity Scores and Severity Weights for Multiple Toxicities in Dose-Finding Oncology Trials

Nathaniel S. O'Connell
Medical University of South Carolina

Elizabeth Garrett-Mayer
Medical University of South Carolina

A latent variable approach to elicit continuous toxicity scores and severity weights for multiple toxicities in dose-finding oncology trials

Nathaniel S. O'Connell*[1] and Elizabeth Garrett-Mayer†[1,2]

[1]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA
[2]Hollings Cancer Center, Medical University of South Carolina, Charleston, SC, USA

Abstract

Most dose-finding clinical trials in oncology aim to find the highest dose yielding an acceptable toxicity profile for patients. The conventional dose-finding framework utilizes a binary toxicity endpoint that treats low to moderate toxicities as irrelevant, ignoring potentially harmful combinations of such toxicities. A handful of novel dose-finding methods have been introduced that combine multiple toxicities across varying grades into a composite toxicity severity score. Toxicity scores provide the advantage of accounting for all toxicity information in a patient profile, but calculation of such scores require prior specification of toxicity severity weights to represent the relative toxicity burden each toxicity type of each grade adds to a toxicity profile if observed. Elicitation of severity weights generally rely on subjective specification, and resulting continuous scores may be confusing in clinical settings. In a statistical framework, we propose a novel method of estimating toxicity weights via a cumulative logit model, assuming there to be a latent continuous toxicity score characterized by the set of observed toxicity types and grades a patient exhibits. Toxicity scores are directly associated with an ordinal outcome assigned to toxicity profiles by clinicians, which correspond to simple dose escalation decisions. The toxicity score elicitation method (TSEM) produces an accurate toxicity scoring system through evaluation of a balanced subset of toxicity profiles in terms of severity, and we present an adaptive weight finding algorithm to facilitate this. This approach bridges the gap between relating continuous toxicity scores to clinically logical ordinal outcomes akin to traditional toxicity grades, and provides an objective method for determining toxicity weights and scores.

Keywords: Phase I; Clinical Trials; Dose-Finding; RP2D; Maximum Tolerated Dose; MTD; Oncology; Continual Reassessment Method; Toxicity Profile; CRM; Toxicity Scores; Toxicity Weights; TSEM

*oconneln@musc.edu
†garrettm@musc.edu

1 Introduction

Most dose-finding (phase I) clinical trials in oncology aim to find the maximum tolerated dose (MTD), also known as the recommended phase II dose (RP2D), for an experimental treatment based on some pre-specified acceptable level of toxicity experienced by patients. Under the traditional phase I framework, MTD determination has been designed around a bivariate toxicity outcome, where each patient has either one or more dose limiting toxicities (DLT) or no DLTs within a defined time frame (e.g. one cycle of treatment). The MTD is the highest dose falling below or within a range of an acceptable percentage of patients expected to experience a DLT. Conventionally, adverse events follow an ordinal measure ranging from 1-5 in terms of increasing severity, defined by the Common Toxicity Criteria for Adverse Events (CTCAEv4.0) published by National Cancer Institute. DLTs in the traditional binary framework are usually classified by an observed grade 3 or 4 toxicity depending on toxicity type, and any single toxicity type at an unacceptable grade constitutes a DLT for a patient.

The binary toxicity endpoint is utilized by almost all modern model based dose-finding methods, such as the Continual Reassessment Method (CRM) (O'Quigley, Pepe, and Fisher, 1990), and its variants. However, this simplification ignores low to moderate grade toxicities, simply treating all non-DLT toxicities as irrelevant. For example, if a grade 4 toxicity is considered a DLT for a particular toxicity type, then a grade 3 toxicity is treated equally as no toxicity at all. Furthermore, potentially harmful combinations of several low-moderate grade toxicities occurring together are ignored.

In the past decade, a handful of model based dose-finding methods have been proposed around a composite toxicity score incorporating multiple toxicities of varying grades to address these limitations. Bekele and Thall (2004) first proposed a total toxicity burden score (TTB), a summary measure of multiple toxicities over varying grades. A vector of toxicities and grades are mapped to relative severity weights, and the sum of severity weights for observed toxicities yield the TTB. A target TTB is elicited in close collaboration with clinicians, and is targeted similarly as the MTD is in conventional designs. Yuan, Chappell,

and Bailey (2007) propose a modified CRM for modeling quasi-continuous toxicity scores using a quasi-Bernoulli-likelihood and normalized toxicity scores $\in \{0, 1\}$. Severity weights are mapped to the ordinal grading system like in the TTB design of Bekele and Thall (2004), but fail to account for multiple toxicities observed in a single patient, and as follows, fail to differentiate between toxicity types with regards to their relative contribution to the toxicity profile. Chen et al. (2010) make note of this, and propose an equivalent toxicity (ET) score and dose-finding method, mapping adjusted grades akin to severity weights to toxicities.

Lee, Cheng, and Cheung (2011) present a modified CRM design that can handle both quasi-continuous or ordinal toxicity outcomes. Incorporating toxicity constraints into the framework of the dose-finding model, they address the issue of target toxicity scores not corresponding to a reliable and grounded definition of an MTD. Most recently, Ezzalfani et al. (2013) proposed a CRM extension similar to Yuan et al. (2007) utilizing a quasi-Bernoulli-likelihood framework, but with a differing toxicity score summary measure referred to as the total toxicity profile (TTP). Like Bekele and Thall (2004) and Yuan et al. (2007), severity weights are mapped to toxicities by grade, but the TTP is calculated as the Euclidean norm of a mapped weight matrix to a toxicity profile.

Each proposed method shares a common similarity; the toxicity score is calculated as a function of severity weights mapped to each toxicity type and grade. In all but one design, specification of toxicity weights are subjectively determined, requiring close collaboration between clinical investigators and statisticians (Bekele and Thall, 2004; Ezzalfani et al., 2013). Clinical investigators are generally asked to assign subjective severity weights to individual toxicities across grades, usually without context of other potential occurring toxicities. This process can prove to be a daunting task, especially as the expected number of individual toxicities to be considered grows. Poorly defined weight schemes can lead to truly acceptable doses being deemed unacceptable, or worse, overly toxic scores being deemed acceptable (Iasonos, Zohar, and O'Quigley, 2011). Further, after a weighting scheme is defined, a target toxicity score must also be determined, which often requires multiple re-evaluations of

weights to alleviate inconsistencies between acceptability of complete profiles and the target score (Bekele and Thall, 2004; Lee et al., 2012), and as addressed by Lee et al. (2011), "may not correspond to a clinically sound definition of the MTD".

Attention to determining toxicity weights and/or scores is scarce in the current literature. At the time of this writing, Lee et al. (2012) propose the only method of toxicity weight elicitation not solely based on subjective specification. In a statistical framework, they propose a simple method of eliciting weights via linear regression, based on evaluation of historical data and hypothetical patient profiles. This process requires clinical investigators to score complete toxicity profiles on a continuous scale, denoted as a toxicity burden score (TBS), around an acceptability threshold. Each toxicity type over varying grades is treated as a binary predictor, and weights are reflected by the regression coefficients estimated in the linear regression model. Like Bekele and Thall (2004), the TBS is simply the sum of weights mapped to observed toxicities.

This paper builds on the principal idea introduced by Lee et al. (2012), using a statistical modelling approach to elicit toxicity weights. However, rather than rating toxicity profiles on an arbitrary continuous scale around a defined acceptability threshold, toxicity profiles are evaluated by clinical investigators on an ordinal scale corresponding to simple dose-escalation decision rules. It does not necessarily require historical patient data, but only a set of hypothesized toxicities deemed likely to occur. The proposed method provides a simple and relatively quick way to elicit toxicity weights, which at the very least, may act as a strong starting point for defining a toxicity weight matrix. Furthermore, there exists a disconnect between continuous toxicity scores and the traditional ordinal grading system, creating issues with the interpretation of scores by clinical investigators and their ability to relate outcomes to patients (Iasonos et al., 2011; Ezzalfani et al., 2013). We address this concern, and the proposed method bridges this gap by relating continuous toxicity scores to an ordinal scale directly related to dose escalation decisions: a seemingly natural extension to the traditional toxicity-grade classification system.

2 Use of Continuous Toxicity scores in Dose Finding Models

The aim of this paper is not to develop a model for continuous toxicity scores, but rather to propose an approach for elicitation of the scores themselves to be implemented. We summarize for context, in a general sense, an example dose-finding design utilizing continuous toxicity scores. For explicit details on such dose-finding designs and their operating characteristics, see Bekele and Thall (2004), Lee et al. (2011), and Ezzalfani et al. (2013).

The example presented follows the approach of Ezzalfani et al. (2013) and Yuan et al. (2007). The process begins with defining severity weights that map to each possible toxicity type l, $l = 1, ...L$, for each grade g, $g = 1, ..., G$, which may be represented as an $L \times G$ weight matrix $\mathbf{W}$. A continuous toxicity score, θ, is then calculated as a function of the weights mapped to observed toxicities of varying grades in a patient. Formally, let θ_j be a continuous toxicity score for the unique toxicity profile j exhibited by patient t. Next, define a normalized toxicity score to be $\theta_j^* \in [0, 1]$, where $\theta_j^* = \theta_j/\theta_{max}$, is considered as a fractional event. θ_j^* is then modeled via a quasi-Bernoulli likelihood (Papke and Wooldridge, 1996) and a modified quasi-CRM (Ezzalfani et al., 2013; Yuan et al., 2007). Assume for an experimental agent in a phase 1 trial, there are m discrete dose levels, $d_1, ..., d_m$, corresponding with a functional dose-toxicity curve monotonically increasing in d_m and η, $E(\theta^*|d_m) = \psi(d_m, \eta)$. Consider the most recent patient t treated at d_m experiences toxicity score θ_j^*, then θ^* is modeled through a quasi-Bernoulli likelihood, and patient $t's$ contribution is,

$$\Psi(\theta_t^*|\eta) = E(\theta^*|d_m)^{\theta_j}(1 - E(\theta^*|d_m)^{1-\theta_j}$$

with the likelihood updated by,

$$\mathcal{L}_t = \mathcal{L}_{t-1}\Psi(\theta_t|\eta)$$

The parameter, η, can then be estimated via MLE in a likelihood framework, or posterior

mean in a Bayesian framework with the posterior density,

$$\pi_t(\eta) = \frac{\pi_{t-1}(\eta)\Psi(\theta_t|\eta)}{\int \pi_{t-1}(\eta)\Psi(\theta_t|\eta)d\eta}$$

The dose assigned to the next patient $t+1$, after t subjects have been evaluated, is the dose level minimizing $\Delta(\psi(d_m, \widehat{\eta}_t), \theta^*_{DLT})$, where $\Delta(v, w)$ is some measure of distance (e.g. absolute difference), and θ^*_{DLT} is the target score, which can be thought of as the maximum tolerated score on the toxicity score continuum. This process continues until the maximum number of patients have been treated or pre-specified stopping rule is reached.

3 Methodology

From the dose finding model, the proposed method focuses on eliciting the toxicity weight matrix, $\mathbf{W}$, corresponding toxicity scores, θ, and a target score, θ_{DLT}.

3.1 Defining Severity Levels

We assume each combination of toxicities over varying grades corresponds to a continuous latent toxicity score. The toxicity-grade weight elicitation process begins with generating a subset of toxicity profiles to be evaluated by clinical investigators, such that hypothetical toxicity profiles are assigned an ordinal severity level (SL) by the rater. Each toxicity profile represents a complete hypothetical toxicity profile for a patient exhibiting a grade g toxicity, for each toxicity type l. Appealing to CTCAEv4.0, adverse events are defined on an ordinal scale from 1-5 (i.e. $G{=}5$), although only profiles with grades 1-4 are evaluated (grade 5 toxicity corresponding to patient death requires a special set of decision rules as deaths are generally considered unacceptable and usually implies stoppage).

For each toxicity profile, clinical investigators rank each profile in terms of its acceptability based on the question, "If a single patient presented the given toxicity profile, rank the toxicity profile from 1-5 in terms of increasing severity, corresponding to the follow decision rules

for the next enrolled patient": (1) Acceptable, escalate dose by 2+ levels, (2) Acceptable, escalate dose by 1 level, (3) Acceptable, repeat at current dose level, (4) Not Acceptable, de-escalate dose by 1 level, and (5) Not Acceptable, de-escalate dose by 2+ levels.

Given a set of hypothetical profiles and their *SL* scores, *SLs* can then be used to estimate a toxicity scoring scheme, outlined in the following section.

3.2 The Toxicity Score Elicitation Method (TSEM)

To elicit toxicity weights and corresponding scores, we begin by defining J possible toxicity profiles, where $J = (G+1)^L$. We let SL_j denote the ordinal *SL*, k, $k = 1, ...K$, assigned to profile j, $j = 1, ..., J$. Consider the included toxicity types, $\mathbf{X} = X_1, ..., X_L$, to be the set of predictors, each taking on the value g corresponding to the observed ordinal toxicity grade. SL_j are modeled through a proportional odds model,

$$\text{logit}[p(SL_j \leq k \mid x)] = \alpha_k - \boldsymbol{\beta'}\mathbf{X} \tag{1}$$

where α_k represents the k^{th} intercept or cutpoint for $k = 1, ..., K-1$. Note, the *SLs* proposed range from 1 to 5, thus $K - 1 = 4$ within the presented framework. $\boldsymbol{\beta'}$ is the parameter vector for the included set of toxicity types modeled as predictors, with $\boldsymbol{\beta'_l}$ being the constant parameter vector across cumulative logits pertaining to each toxicity type l. From this, define the estimated toxicity-grade weights for each toxicity type l of grade g to be, $\boldsymbol{\beta'_l}\mathbf{X_l} = w_{l,g}$, where each weight, $w_{l,g}$, is mapped to an $l \times g$ toxicity weight matrix $\mathbf{W}$.

Next, let $\theta = \{\theta_1, \theta_2, ..., \theta_J\}$ denote the true underlying continuous toxicity score of profile j, and $\hat{\theta} = \{\hat{\theta_1}, ..., \hat{\theta_J}\}$ be the estimated toxicity scores, such that for each toxicity profile j, $\hat{\theta_j} = \sum_{l=1}^{L} \beta_l X_{lj}$, with the CDF, $P(\theta_j \leq a_k \mid \mathbf{X}) = \Phi(\alpha_k - \hat{\theta_j})$, where $\Phi(\cdot)$ is a valid CDF. For a given toxicity profile satisfying the inequality, $SL_j = k - 1$ if $\alpha_{k-1} \leq \theta_j < \alpha_k$, the estimated

probability of falling into a given SL is,

$$\begin{aligned} P(SL_j = k|\mathbf{X}) &= P(SL_j \leq k \mid \mathbf{X}) - P(SL_j \leq k-1 \mid \mathbf{X}) \\ &= P(\hat{\theta}_j \leq \alpha_k \mid \mathbf{X}) - P(\hat{\theta}_j \leq \alpha_{k-1} \mid \mathbf{X}) \\ &= \frac{\exp(\alpha_k - \hat{\theta}_j)}{1+\exp(\alpha_k - \hat{\theta}_j)} - \frac{\exp(\alpha_{k-1} - \hat{\theta}_j)}{1+\exp(\alpha_{k-1} - \hat{\theta}_j)} \end{aligned} \tag{2}$$

That is, each estimated toxicity score relates probabilistically to an ordinal SL. We exploit the latent variable characteristics of the cumulative-logit to define a latent toxicity score as a function of observed toxicity types and corresponding grades, which then translate back to a more intuitive ordinal outcome. To clarify, each toxicity profile j has both an estimated SL_j and severity score θ_j. θ_j allows for a more refined, estimated ordering of toxicity profiles, while corresponding $SL's$ give a clinically interpretable meaning for oncologists and patients.

Given a weight matrix $\mathbf{W}$, $\hat{\theta}_j$ may be estimated for any possible toxicity profile. With SLs representing a dichotomy between acceptable and unacceptable profiles along the ordinal scale (i.e. between $SL = 3$ and $SL = 4$), an acceptable target score (or range of scores) need not be separately determined, but instead manifest around cutpoint parameters. Denote this acceptability threshold to be α_{tox}, and a target score to be θ_{DLT}, then dose-finding may continue considering a toxicity profile to be an acceptable response response when $\hat{\theta}_j < \theta_{DLT}$, and unacceptable when $\hat{\theta}_j \geq \theta_{DLT}$. An obvious choice for θ_{DLT} may be defined as,

$$\theta_{DLT} = \max_{\theta_j}[P(\theta_j < \alpha_{tox}) > P(\theta_j \geq \alpha_{tox})] = \max_{\theta_j}(\theta_j < \alpha_{tox}) \tag{3}$$

That is, we target the highest score that is predicted to result in $SL = 3$. A second choice,

$$\theta_{DLT} = \overline{\theta} \quad \forall \quad (\alpha_{tox-1} \leq \theta_j < \alpha_{tox}) \tag{4}$$

yields a more conservative target. That is, θ_{DLT} is the mean toxicity score for all toxicity scores with corresponding $SL = 3$, the severity level defined as "repeat at current dose level."

These are certainly not the only choice for target scores, but are two simple possibilities making use of the TSEM. Previous designs have proposed selecting a target score around dose escalation decisions (Bekele and Thall, 2004; Yuan et al., 2007; Ezzalfani et al., 2013), however, like elicitation of the weights themselves, the process can be burdensome and require multiple reevaluations of the weighting scheme to begin with. The TSEM naturally introduces a target score (or range of scores) to target based off of dose escalation decisions, which at the very least, eases the burden of deciding upon θ_{DLT}.

3.3 Multiple Raters

A single clinical investigator never has sole input on designing a trial, thus it follows we should not rely on only a single person's opinion of ranked profiles. In the case of multiple raters, we propose each clinical investigator independently ranks profiles, and results across investigators are combined together and analyzed. This greatly increases the sample size of evaluated profiles, and hence the reliability of w_{lg} estimates. Some investigators will be more liberal in what they consider an acceptable profile, and some more conservative. To account for this, we propose combining all data and controlling for within-observer bias with random intercepts for each rater. The model then becomes,

$$\text{logit}[p(SL_{i,j} \leq k \mid x)] = \alpha_k + \gamma_i - \beta'\mathbf{X} \tag{5}$$

Where, γ_i is the random intercept for the $i'th$ rater. It then follows that α now act as the overall set of cutpoints after adjusting for intra-rater bias, while individual level thresholds may be defined by, $\alpha_{i,k} = \alpha_k + \gamma_i$, with the estimated weight matrix, $\mathbf{W}$, constant across raters. $\mathbf{W}$ is then used to estimate $\hat{\theta}_j$, for any profile j, just as in the single rater scenario.

With multiple raters for comparison, we have a frame of reference to assess reliability by evaluation of intra- and inter-rater variability. It further provides the ability to identify individual toxicity types with greater uncertainty around the true severity relationship with

increasing grades. Such instances would be of greater interest and importance when refining the weights. Exploration of such characteristics extend beyond the scope of this paper given length considerations, but will be of interest in subsequent research.

3.4 Model Characteristics and Assumptions

The TSEM makes the following assumptions: (1) Within a toxicity type, increasing toxicity grades should be non-decreasing in severity. Failure to adhere to this assumption will lead to nonsensical results, and if such results appear, SL assignments should be revisited. (2) The TSEM assumes individual toxicities act independently in their contribution to the overall toxicity a patient experiences, i.e. toxicity scores are an additive function of weights.

In exploration of various working models, we propose a quadratic trend, striking a balance between parsimony and flexibility, with weights defined individually by, $w_{lg} = \beta_{l,linear} X_l + \beta_{l,quadratic} X_l^2$. Allowing for a quadratic increase in toxicity weights makes clinical sense, such that high grade toxicities (i.e. grade 3 or 4) are likely to be much more severe than lower grades. A quadratic trend will allow such relationships to be more accurately estimated, but cases of truly linear relationships will be reflected as well.

Constraints are not imposed on parameter estimation, such that coefficients may be negative, which in turn can lead to negative toxicity weights at low grades. For example, if for X_l at grade g^*, $\beta_{l,linear} X_l + \beta_{l,quadratic} X_l^2 < 0$, then $w_{l,g^*} < 0$. Given toxicity scores are an assumed additive function of severity weights, a low grade toxicity with a negative weight would result in a lower score when combined with a high grade toxicity than the high grade toxicity would have alone. These results are clinically illogical, therefore negative weights are adjusted to 0 after estimation (i.e. contribute nothing to the overall toxicity profile). Such instances may happen particularly in cases when additivity is violated; the best fitting model mathematically may be one with negative parameters for low grades, compensating in a sense, for an unknown interaction. While this negates the ability to use information of low grade toxicities to predict higher grades, there will never be any less information than

in the traditional binary response CRM, given that for individual grades of a single toxicity, $\hat{w}_{l,g} \geq \theta_{DLT}$ when $w_{l,g}$ is truly too toxic.

When additivity is not severely violated, this should be a non-issue. The number of parameters estimated, $2(K-1)+2L$, may lead to large variability estimates. However, the objective is not one of inference, but rather estimation and prediction based on an assumed relationship of weights. Assuming additivity and a quadratic trend, if the spectrum of SLs are well represented in the subset of evaluated profiles, the TSEM provides unbiased estimates of $w_{l,g}$ through $\hat{\boldsymbol{\beta}}_{\mathbf{l}}'\mathbf{X_l}$,

Theorem 1. *Let $w_{l,g}$ be 'true' toxicity-grade weights from $\mathbf{W}$, estimated by $\hat{w}_{l,g}$ through $\hat{\boldsymbol{\beta}}_{\mathbf{l}}'X_l$ as defined in section 3.2. If θ is truly an additive function of weights from $\mathbf{W}$, and $w_{l,g}$ can be represented quadratically within each type l, then $\hat{w}_{l,g}$ is a proportionally unbiased estimate of $w_{l,g}$. That is, for some fixed constant C:*

$$E(C\hat{w}_{l,g}) - w_{l,g} = 0 \ \forall \ l, g$$

Alternatively, let $w^*_{l,g}$ be an additional 'true' weight arising from the same $\mathbf{W}$ and estimated by $\hat{w}^*_{l,g}$, then:

$$E(\frac{\hat{w}_{l,g}}{\hat{w}^*_{l,g}}) - \frac{w_{l,g}}{w^*_{l,g}} = 0 \ \forall \ l, g$$

Under the given assumptions, the proof is straightforward appealing to well-known statistical principles, demonstrated in the appendix.

Given that SLs are not uniformly distributed across profiles, but heavily skewed towards more severe profiles, we cannot rely on random selection. Efficient estimation is best achieved through evaluation of the smallest subset with the spectrum of SLs evenly represented. Suppose a subset n profiles are evaluated from the complete set of size J, let n_k be the number of evaluated profiles with true $SL = k$, such that $\sum_{k=1}^{5} n_k = n$. Then $\mathbf{W}$ and θ will be most efficiently estimated when, $(n_1 \approx n_2 \approx ... \approx n_5)$, which we hereby refer to as the 'SL equivalence constraint'. We propose an adaptive weight finding algorithm to ensure the

set of evaluated profiles are representative of the SL range.

3.5 Weight Finding Algorithm

The weight finding algorithm (WFA) hinges on the assumption that user rater scores reflect the true scores. Let $\mathbf{X} = \{X_1, X_2, ..., X_J\}$ be the complete set of possible toxicity profiles, and let $\mathbf{SL} = \{SL_1, SL_2, ..., SL_J\}$ be the set of true SLs for $\mathbf{X}$. We now introduce the notion of observed and unobserved profiles, where an observed profile j has an assigned SL_j as determined by the rater, and an unobserved profile has not been assigned a SL by the rater. We combine a profile, X_j and its severity level, SL_j, to form the complete profile, $Y_j = (X_j, SL_j)$, which is either be observed, $Y_j^{(obs)}$, or unobserved $Y_j^{(unobs)}$. It then follows $\mathbf{Y} = \{Y_1, Y_2, ..., Y_J\}$ is the entire set of complete profiles. Let $n^{(obs)}$ denote the length of $\mathbf{Y^{(obs)}}$ observed profiles drawn from $\mathbf{Y}$, and let $n_k^{(obs)}$ denote the length of subset $Y_k^{(obs)}$, where $Y_k^{(obs)}$ is the subset of $\mathbf{Y^{(obs)}}$ where $SL = k$; the goal is to ensure, $(n_1^{(obs)} \approx n_2^{(obs)} \approx ... \approx n_5^{(obs)})$ in the evaluated subset of the TSEM.

For parameters to be estimable, the model requires at least one outcome for each SL to be observed (i.e. $n_k^{(obs)} \geq 1 \forall k$). Early on in the weight elicitation process, parameter estimation will be unstable. To stabilize it, we introduce pseudo data, that is, a prior set of complete toxicity profiles with assigned SLs, denoted $\mathbf{Y^{(p)}}$, of length $n^{(p)}$. Pseudo data represent a subset of toxicity profiles mimicking observed data; however, pseudo data are only used to define a prior working model with estimable parameters when the set of observed data is relatively small. In a similar manner to the observed data, we define $\mathbf{Y^{(p)}} = \{X^{(p)}, SL^{(p)}\}$ to be a set of complete pseudo profiles arising from a pre-specified pseudo weight matrix, $\mathbf{W^{(p)}}$. Note, $n^{(p)}$ is not of length J, but rather just large enough to ensure an estimable set parameters when there is no observed data yet recorded. $SL^{(p)}$, the SLs assigned to $X^{(p)}$, may be assigned based on some arbitrarily defined set of decision rules, so long as they are consistent, satisfy the 'SL equivalence constraint' and assumptions defined in section 3.4.

With a set of pseudo-profiles defined, the WFA begins by drawing a randomly selected

toxicity profile, X_1 from $\mathbf{X}$, and assigning it a SL score, SL_1, thus giving the first complete observed profile, $Y_1^{(obs)}$. With the first profile drawn, the set of un-evaluated profiles is defined by $\mathbf{Y^{(unobs)}} = \mathbf{Y} - \mathbf{Y^{(obs)}}$. We combine pseudo data and true data profiles into a 'complete' data set, letting, $X^{(c)} = \{X^{(obs)}, X^{(p)}\}$ and $SL^{(c)} = \{SL^{(obs)}, SL^{(p)}\}$, which form $\mathbf{Y^{(c)}}$. The complete multinomial likelihood is then given by,

$$L(\alpha, \beta) = \prod_{j=1}^{n^{(c)}} \prod_{k=1}^{5} (p_{jk}(\alpha, \beta))^{SL_{jk}^{(c)}} \tag{6}$$

Where p_{jk} is the probability of severity level k for toxicity profile j and $n^{(c)} = n^{(p)} + n^{(obs)}$. From this, the TSEM is fit as,

$$\text{logit}[p(\mathbf{SL^{(c)}} \leq k | \mathbf{X^{(c)}})] = \alpha_k - \boldsymbol{\beta'}\mathbf{X} \tag{7}$$

Let $\mathbf{Y^{(pred)}}$ contain the predicted set of SLs over the entire set of profiles predicted by equation 7. Given our observed data after j evaluated profiles, $\mathbf{Y^{(obs)}} = \{Y_1, ..., Y_j\}$, each successively drawn profile, X_{j+1} is selected randomly from the subset $Y_k^{(unobs)}$ where

$$\{Y_k^{(unobs)} \leq Y_{k^*}^{(unobs)} \, \forall \, k \neq k^*\}$$

Which is to say, X_{j+1} is drawn such that,

$$\{SL_{j+1}^{(pred)} = k \mid n_k \leq n_{k^*} \, \forall \, k \neq k^*\}$$

To clarify, the $j+1$ profile is adaptively selected such that it's predicted SL, $SL_{j+1}^{(pred)}$, is under-represented in the already observed set of profiles. For example, if 3 profiles have been evaluated, and $SL^{(obs)} = \{3, 1, 4\}$, the next profile evaluated, X_4, will be selected from the subset of $Y^{(unobs)}$ where $SL^{(pred)} = \{1, 2\}$. As $n^{(obs)}$ grows, $\mathbf{Y^{(p)}}$ are down-weighted in the model such to allow profile selection and weight elicitation to be driven by $\mathbf{Y^{(obs)}}$. The

proposed weighting scheme is defined by the sample size, $n^{(obs)}$, at which $\mathbf{Y^{(p)}} = Y_j^{(obs)}$ in terms of overall contribution to the fitted model. In other words, the sample size at which the weighted contribution from the entire set of pseudo profiles equals that of a single true profile. Let this point be defined by τ, and define $\omega^{(p)}$ to be the weight of each $Y_j^{(p)}$, then after $n^{(obs)}$ observed profiles,

$$\omega^{(p)} = \frac{\tau \times (1/n^{(p)})}{n^{(obs)}} \tag{8}$$

Evaluation continues in this fashion until a sufficient number of observed profiles have been evaluated (e.g. $n_k > 10 \quad \forall \quad k$). Further stopping criteria could also be set in conjunction with a minimum size, such as MSPE, % of toxicities correctly predicted over $Y^{(obs)}$, etc.

The idea behind using pseudo-data is ideologically Bayesian in the sense that we start with a prior model and continuously update the predictive model as true information becomes available. With only a few true observations early on (i.e. when $n^{(obs)}$ is small), parameters in the model must have a priori specification to be able to update the model. With the elicitation process driven by clinical investigators ranking profiles in real time, we strive for instantaneous results. Rather than relying on MCMC, we generate pseudo-data and rely on maximum likelihood estimation because it eases the computational burden and time it takes to update the model. Weighting $\mathbf{Y^{(p)}}$ inversely as a function of increasing $n^{(obs)}$ ensures that as more true information becomes available, parameter estimation (and hence w_{lg} estimation) is determined primarily by $\mathbf{Y^{(obs)}}$. Just as one can specify informative and non-informative priors in a Bayesian context, we can influence the 'informativeness' of the pseudo data by increasing or decreasing τ and/or $n^{(P)}$.

4 Simulation

Simulations were performed in R 3.1.1 (R Core Team, 2014), with use of the 'ordinal' and 'sampling' packages (Christensen, 2015; Till and Matei, 2013). We first evaluate the operating characteristics of the TSEM over several 'true' toxicity profiles of varying toxicity types l, across a range of $n_{(K)}$. We then demonstrate results of the WFA, and evaluate it in terms

of its ability to select profiles in accordance with the 'SL equivalence constraint'.

4.1 Simulation: TSEM Operating Characteristics

For a given number of toxicities, L, we generate a complete set of J 'true' toxicity profiles. We assume these profiles to reflect the set of true beliefs for an individual rater, if that rater could assign specific scores to each and every profile. We begin by defining a 'true' toxicity weight matrix, $\mathbf{W}$. All possible J toxicity profiles are enumerated, and by mapping each toxicity profile's observed toxicity type and grade to $w_{l,g}$ through $\mathbf{W}$, θ_j is calculated by $\theta_j = \sum_{l=1}^{L} w_{l,g} I(X_l = g)$, where $I(X_l = g) = 1$, if there is an observed grade g toxicity for toxicity type l, and 0 otherwise. The set of profiles are sorted by θ_j, ordered from least to most severe. $K-1$ threshold points are established along the continuum of θ_j, establishing inequalities used to define SL_j; this establishes the 'true' complete set $\mathbf{Y}$. These cutpoints are selected based the defined $\mathbf{W}$ and what would be clinically logical for the given scenario.

With $\mathbf{Y}$ established, we test the TSEM by selecting a random subset of profiles from $\mathbf{Y}$ under the SL equivalence constraint and estimate the weights through the TSEM. The scale of weights will differ for each scenario, but the ratio of weights relative to one another remain comparable, so we present results on a normalized scale by defining weights as, $w^*_{l,g} = \frac{w_{l,g}}{max(w_{l,g})}$, so that $w^*_{l,g} \in [0,1]$. Similarly, SL threshold cutpoints, α_k are adjusted to the same scale by, $\alpha^*_k = \frac{\alpha_k}{max(w_{l,g})}$. We specify 4 unique toxicity profiles, for each of $L = 3, ..., 7$ toxicity types, and simulate the TSEM across values of $n_k = (8 \;\&\; 15)$. Simulations are performed under two scenarios. The first scenario generates the 'true' set, $\mathbf{Y}$, assuming additive weights in accordance with the assumption made by the TSEM. A second scenario simulates $\mathbf{Y}$ via a non-additive function of weights. Ezzalfani et al. (2013) proposed applying the Euclidean norm to an observed toxicity profile to define the severity score, i.e.

$$\theta_j = \sqrt{\sum_{l=1}^{L} \sum_{g=0}^{4} w^2_{l,g} I(X_l = g)}$$

As a property of Euclidean distance, the triangular inequality naturally penalizes multiple toxicities. We appeal to this property to easily simulate non-additive toxicity scores and test the TSEM when additivity does not explicitly hold.

Results are presented in terms of AIC, mean square predictive error (MSPE), percentage of $SL^{pred} = SL^{true}$, and percent of DLTs correctly predicted (i.e. truly toxic doses and non-toxic doses correctly predicted). In the latter, we define DLT as a profile with predicted $SL > 3$. Stemming from this, we also present the percentage of under-estimated DLTs (i.e truly too toxic profiles with $SL_j^{pred} \leq 3$), and over-estimated DLTs (i.e truly safe profiles with $SL_j^{pred} > 3$). In other words, we evaluate the overall performance of the TSEM by its ability to correctly predict true SLs given the estimated weights.

4.2 Simulation: Weight-Finding Algorithm

We first simulate a true data set as in section 4.1. We next define our pseudo data as outlined in section 3.5 in the same way we generate true data in 4.1, and select a subset in accordance to the 'SL equivalence constraint' to act as $\mathbf{Y^{(p)}}$. $\mathbf{Y^{(p)}}$ is not the entire set of J possible profiles, but a subset just large enough to give estimable parameters when n^{obs} is small. Pseudo data are generated from a vague, $\mathbf{W^{(p)}}$, where each toxicity type follows the same increasing trend of weights across grades, $W_{l,\cdot} = (0.1, 0.3, 0.8, 1.5)$.

Following the WFA outlined in section 4.1, the simulation is straightforward. Randomly draw a single profile from the true data set, combine $\mathbf{Y^{(obs)}}$ and $\mathbf{Y^{(p)}}$ to form $\mathbf{Y^{(c)}}$, fit the TSEM for $\mathbf{Y^{(c)}}$, predict the fit over $\mathbf{Y}$, adaptively draw X_{j+1} from $\mathbf{Y^{(unobs)}}$ in accordance to section 4.1, and reiterate until stopping criteria have been met. We run simulations over 500 iterations for each scenario, and use pre-specified stopping criteria based on $\min(n_k) \forall k$, and $\max(n)$. We evaluate $\min(n_k) = (8 \& 15)$, and for each, set $\max(n) = (\min(n_k) \times k) + 20$. We present results in terms of criteria in section 4.1, as well as the mean and median number of evaluated profiles and the percentage of simulations that reach the maximum.

5 Results

For the sake of brevity, presented results primarily focus on the scenario where $L = 4$ included toxicity types and $n_{(k)} = 8$, i.e. $n = 40$ total profiles evaluated out of $J = 625$ total profiles, although results and discussion extend to $L = 3, 5, 6, \& 7$, and $n_{(K)} = 15$.

5.1 TSEM Simulation Results

True weights, estimated weights, and SL thresholds can be found in table 2, with diagnostic results presented in table 3. We refer to each scenario using the notation, '$L.n$', where L refers to the number of included toxicities in the scenario, and n denotes the scenario number. With an evaluated subset of $n = 40$ (i.e. $n_k = 8$), the TSEM correctly predicts the true SL across $J = 625$ profiles with 83.68% to 91.56% accuracy (profiles 4.4 and 4.2 respectively) when toxicity scores are truly an additive function of weights. The accuracy is slightly lower when the toxicity scores are truly penalized through the EN function (but additivity is assumed), with the TSEM accurately predicting between 79.65% to 85.04% across the entire true set. These differences are reflected in goodness of fit measures, with results from EN-penalized weights resulting in marginally higher AIC and MSPE.

Evaluating profiles with regards to a binary toxicity endpoint on the continuum of scores - too toxic vs acceptable - the TSEM predicts with 95.44% to 97.66% accuracy in a truly additive model, and with 95.19% to 96.12% accuracy under the EN penalized model. As an example from table 2, under scenario 4.2-additive, the TSEM accurately predicts percentage of toxicities correct with 97.66% accuracy, of which 1.63% (10 profiles) are under-estimated (i.e. predicted to be safe when they are truly too toxic), and 0.71% (4 profiles) are over-estimated. Performance is similar across the other 3 scenarios presented, and slightly worse under the EN penalized weights. For instance, with EN penalized true weights in 4.2, 3.64% (23 profiles) are under-estimated as being acceptable when they're truly too toxic.

Despite additivity not holding, the TSEM still leads to a reasonable toxicity scoring

structure. However, with the true EN penalized weight function, weights themselves tend to be underestimated, particularly effecting lower grade toxicities. Comparing true vs. estimate weight profiles in table 2, we see that estimated grade 1 weights are at or close to 0 in almost all cases with a true EN penalized relationship. Similarly, grade 2 weights are generally lower than the true weights, and the differences between low grade toxicity weights within a particular weighting scheme are closer together than in truth. This makes comparison of combinations of low grade-toxicities difficult.

Indistinguishable weights in these circumstances are in part due to the fact that weights are adjusted to 0 when $\beta_l' X_l < 0$. Allowing weights to be less than 0 would provide greater insight to the relationships between low-grade weights, but would lead to nonsensical toxicity scores when combined. For any general interaction that may results in an abundance of negative estimated weights, it would be of interest to adapt the model to account for them. While investigating such possibilities is beyond the scope of the framework we layout in this paper, there are surely several avenues to explore, for example, simply shifting weights and SL thresholds by an additive constant, such to make the lowest estimated weight equal 0.

Generalizing to $L = 3, 5, 6, \& 7$, included toxicity types, the TSEM continues to perform well. As L increases and n_k remains constant, we see a decline in predictive performance - intuitively so, given that we are estimating an additional 2 parameters for each additional toxicity type holding a constant sample size. For example, when $n_k = 8$ in scenarios which most closely resemble 4.4 (the poorest performing model among the $L = 4$ profiles), with truly additive weights, the TSEM predicts true SLs with 93.43% accuracy for scenario 3.4, and with 87.29% accuracy for scenario 7.4. Increasing to $n_k = 15$, predictive accuracy is improved to 91.00%. When $L = 7$, there are 78,125 possible toxicity profiles. Improvement from 87.29% to 91.00% may seem trivial, but this equates to a difference of nearly 3,000 additional profiles being correctly predicted with respect to the SL.

5.2 Weight Finding Algorithm Results

Having demonstrated favorably consistent results when the 'SL equivalence constraint' is upheld, we now focus the WFA's ability to equally allocate profiles on the basis of severity. With $\min(n_k) = 8$, the minimum number of profiles to be evaluated is 40, and we set the maximum at 60. In simulations over 500 iterations, the WFA performs accurately as designed. The mean and median number of profiles evaluated for true SN weights are 44.24 (42) , 43.53 (42), 43.35 (42), and 47.91 (45), for profiles 4.1-4.4 respectively. The number of profiles that reach the maximum of 60 are 12 (2.4%) , 16 (3.2%), 6 (1.2%), and 59 (11.8%). Similarly, for the non-additive EN weights, the mean number of profiles evaluated are 44.16 (42), 44.33 (43), 43.84 (42), and 49.19 (48) for profiles 4.1-4.4 respectively, with the number of profiles that reached the max being, 25 (5%), 5 (1%), 27 (5.4%), and 59 (11.9%).

The EN penalized models perform slightly worse for scenarios 4.1 and 4.3, with profiles reaching the max being double that of the additive model. Conversely, the EN penalized model performs better in scenario 2 and about equally in 4.4. Scenario 4.4 is the poorest performing of the 4 scenarios. From section 5.1, the scenario 4.4 demonstrates the poorest predictive accuracy, thus it follows that the adaptive selection of profiles to be evaluated will also perform poorest given the WFA depends on the models predictive accuracy.

Using the notation, $L.nS = \{n_1, ..., n_5\}$, where S is 'A' for additive or 'EN' for EN penalized, we present the true SL distributions for each scenario, which are as follows: $4.1A = \{23, 32, 56, 74, 440\}$, $4.1EN = \{40, 59, 85, 124, 317\}$, $4.2A = \{13, 25, 50, 79, 458\}$, $4.2EN = \{23, 48, 62, 171, 321\}$, $4.3A = \{20, 27, 56, 89, 433\}$, $4.3EN = \{26, 52, 93, 176, 278\}$, $4.4A = \{15, 34, 66, 52, 458\}$, and $4.4EN = \{24, 79, 42, 167, 313\}$. These make apparent the heavy skew towards more severe profiles, with the number of $SL = 1$ across profiles ranging from 13 to 40. Even with only 13 profiles to select from for $SL = 1$ in $4.2A$, the WFA selects at least 8 profiles from the 13 before hitting the maximum in all but 7 of 500 iterations.

Dependent on predictive accuracy, the WFA leads to an efficient selection of profiles, ensuring the 'SL equivalence constraint'. The use of pseudo data effectively stabilizes the WFA

early, and makes a negligible impact toward the end. In rare cases when $n_k \neq \min(n_k) \, \forall \, k$, it is because early profile selection leads to a 'runaway' selection of highly toxic profiles ($SL = 5$), which may be better controlled for with a more informative $W^{(p)}$.

6 Discussion

Cancer treatments are undergoing a paradigm shift. The introduction of molecularly targeted agents, immunotherapies, and an increased number of combination therapies are leading to increasingly complex patient toxicity profiles. Conventional methods may no longer be optimal in many circumstances, and novel techniques and methods must be developed to meet the demands of the changing landscape of cancer treatments. One such way is through re-defining the way we think about toxicities. Rather than a binary 'yes/no' outcome defined by a single toxicity type, such novel designs account for the fact that many contributing low-grade toxicities, while individually acceptable, may induce an overly-toxic burden on a patient when combined. Bekele and Thall (2004); Yuan et al. (2007); Chen et al. (2010); Lee et al. (2012); Ezzalfani et al. (2013) have proposed the modeling frameworks for handling such composite quasi-continuous toxicity scores, but until this point, little consideration has been given to objectively defining severity weights used to compute these scores. In this paper we have build on the premise of Lee et al. (2011), using a modeling approach to elicit weights, and propose a generalizable method to elicit a toxicity weighting scheme.

We address concerns regarding interpretability (Iasonos et al., 2011; Ezzalfani et al., 2013) through the natural mapping of scores to intuitive ordinal outcomes. Previously, toxicity weights and scores have been based on arbitrary scales, differing from one design to the next. Our method puts all trial designs, regardless of the number or type of included toxicities, on a relatable scale, with ordinal outcomes invariant to the actual scores themselves.

A notable strength of the TSEM is its ability to derive a cohesive and relatively accurate weighting/scoring scheme for 3+ toxicities. Up unto this point, most dose-finding methods for continuous scores have demonstrated their use with toxicity weight matrices with only

2 or 3 included toxicities, generally with the note along the line of "weights are decided upon by expert clinicians." With ≤ 3 included toxicities, there are ≤ 125 possible profiles. With 125 profiles, forming a cohesive weighting scheme and target score may still be feasible through subjectively defining weights by trial and error. With ≥ 4 toxicities, the number of possible profiles grows rapidly, and coming up with a subjective weighting scheme may become overwhelmingly difficult. We have demonstrated that up to 7 toxicity types equating to over 70,000 possible profiles, the TSEM leads to an accurate toxicity scoring scheme.

The TSEM is not without limitations. The quadratic weight trend, while suitable in many clinical situations, may be overly inaccurate when there are 2 or more points of inflection in the toxicity-weight severity curve (as observed in scenario 4.4). For example, if a drastic increase in severity exists between grade 2 and grade 3 within a single toxicity type, but a grade 4 is essentially no different from a grade 3, then following a quadratic trend, the estimated grade 4 weight will likely be grossly over-estimated.

In the same vein, it is unlikely that weights perfectly follow a quadratic trend; as a consequence, individual toxicity weights that borderline the acceptability threshold need to be carefully evaluated. This is particularly apparent when weights are not truly additive, as in the EN penalized scenarios presented. This become especially important to note for truly too toxic grade 3 toxicities not individually exceeding the estimated threshold between $SL = 3$ and $SL = 4$. Take note of scenario 4.4, the 1st and 3rd toxicities are truly too toxic in the true scenario (weights of 0.48 and 0.60, exceeding the threshold of 0.45). However, when the weights are truly EN penalized, no grade 3 toxicity is individually estimated to be too toxic. Such instances may give merit to selecting α_{tox} as defined by equation 4.

These possible occurrences make apparent that the scoring scheme given by the TSEM should not be taken blindly. Evaluation and refinement should be expected. For example, it's easy to subjectively evaluate each individual toxicity (particularly grade 3s and 4s) in terms of if they are too toxic or not, and then re-evaluate the weighting scheme with regards to that. If individual grade 3's should be too toxic, but estimated TSEM weights do not

reflect it, an easy solution could be to simply increase them so they equal or exceed α_{tox}.

A final consideration is to address is the possibility of interactions (i.e. non-additivity of weights). Two weight functions are proposed in the literature, the additive and the EN penalized (Ezzalfani et al., 2013), however, there exists potential for several more complex interactions. Accounting for these prove to be a difficult task; modeling interactions would often lead to unidentifiable parameters. One option for future consideration may be exploration of various post-hoc adjustments. A second possible solution could be to introduce binary indicator variables rather than modeling interactions directly. Bekele and Thall (2004) account for 5 separate toxicities applied to a Sarcoma trial in their proposed method, but note that the severity of the toxicity, 'myelosupression", is dependent on the presence of a fever. In cases such as this, an additional binary indicator could be used, simply adding additional weight in the calculation of scores when a fever is present. This idea could be extended to include additional weight indicators for specific toxicity combinations. Such adaptations would decrease the generalizability, but could improve weight/score estimation, and these ideas provide avenues for future work.

We conclude with noting that the TSEM is not meant to give a perfect weighting scheme, but is designed as an easily implementable method to provide a foundation of weights and scores to be applied in a trial. The TSEM results should always be checked among statisticians and clinicians, and refinement will likely be necessary. But when faced with considering 3 or more toxicities to be included, the TSEM provides an easy way to develop a consistent and relatively accurate crude weighting scheme to build from, with natural target scores and clinically interpretable outcomes. We have laid out the base statistical framework for the TSEM, however, its implementation will be driven primarily by clinical investigators. In light of this consideration, we have developed an app through R-Shiny (Chang et al., 2015) implementing the TSEM, which provides real-time updating of toxicity weights and acceptability thresholds. At the time of this writing, the app is being refined and is for demonstrational purposes; it may be provided by the authors upon request.

References

Bekele, B. N. and Thall, P. F. (2004). Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *Journal of the American Statistical Association* **99,** 26–35.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2015). *shiny: Web Application Framework for R.* R package version 0.12.0.

Chen, Z., Krailo, M. D., Azen, S. P., and Tighiouart, M. (2010). A novel toxicity scoring system treating toxicity response as a quasi-continuous variable in phase i clinical trials. *Contemporary clinical trials* **31,** 473–82.

Christensen, R. H. B. (2015). ordinal—regression models for ordinal data. R package version 2015.1-21. http://www.cran.r-project.org/package=ordinal/.

Ezzalfani, M., Zohar, S., Qin, R., Mandrekar, S. J., and Deley, M.-C. L. (2013). Dose-finding designs using a novel quasi-continuous endpoint for multiple toxicities. *Statistics in medicine* **32,** 2728–46.

Iasonos, A., Zohar, S., and O'Quigley, J. (2011). Incorporating lower grade toxicity information into dose finding designs. *Clinical trials (London, England)* **8,** 370–9.

Lee, S. M., Cheng, B., and Cheung, Y. K. (2011). Continual reassessment method with multiple toxicity constraints. *Biostatistics (Oxford, England)* **12,** 386–98.

Lee, S. M., Hershman, D. L., Martin, P., Leonard, J. P., and Cheung, Y. K. (2012). Toxicity burden score: a novel approach to summarize multiple toxic effects. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO* **23,** 537–41.

O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics* **46,** pp. 33–48.

Papke, L. E. and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* **11,** 619–632.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Till, Y. and Matei, A. (2013). *sampling: Survey Sampling*. R package version 2.6.

Yuan, Z., Chappell, R., and Bailey, H. (2007). The continual reassessment method for multiple toxicity grades: a Bayesian quasi-likelihood approach. *Biometrics* **63,** 173–9.

Appendix

Proof of 1. Let θ denote a true known continuous toxicity score. Let $\{\theta \in \mathbf{\Theta} | \mathbf{\Theta} \in \mathbb{R}\}$, be an additive function of toxicity weights without interaction, such that

$$\theta = \sum_{l=1}^{L} \sum_{g=0}^{4} w_{l,g} I(\max[X(l) = g])$$

For a single observed toxicity, l^*, of grade g, it follows $\{\theta = w_{l^*,g} | w_{l,g} = 0 \, \forall \, l \neq l^*\}$. For a fixed toxicity type l, let $w_{l,g}$ be quadratically increasing in $g \ni \theta$ is represented by the linear model, $\theta = \lambda_0 + \lambda' \mathbf{X}$, where the set of predictors, $\mathbf{X}$, represents each toxicity type l, taking on values g, $g = 0, ..., 4$, and

$$\lambda' \mathbf{X} = \lambda_{1.lin} X_1 + \lambda_{1.quad}(X_1^2) + ... + \lambda_{L.lin}(X_L) + \lambda_{L.quad}(X_L^2)$$

Each weight is uniquely given by, $w_{l,g} = \lambda_{l.lin} X_l + \lambda_{l.quad}(X_l^2)$. Estimation of $\hat{\lambda}$ yields $\hat{w}_{l,g}$ through $\hat{\lambda}' X$. Constraining $\lambda_0 = 0$, which is clinically logical given $\theta = 0$ when $\sum_{l=1}^{L} \sum_{g=0}^{4} w_{l,g} I(\max[X_l = g]) = 0$, it follows that for some fixed $\mathbf{X}$,

$$E(\hat{\lambda}) = \lambda \implies E(\hat{\lambda}' X) = E(\hat{w}_{l,g}) = \lambda' \mathbf{X} = w_{l,g}$$

and thus for each toxicity type l of grade g, $E(\hat{w}_{l,g}) - w_{l,g} = 0$.

However, θ is unknown and unobserved, and is estimated by $\boldsymbol{\beta'}\mathbf{X}$ through equation 1. Let θ^* be the estimate for θ manifesting on the ordinal scale, $SL = 1, ..., 5$. Consider the cumulative logit model motivated by the latent variable, defined by,

$$\text{logit}[p(SL \leq k \mid x)] = \alpha_k - \boldsymbol{\beta'}\mathbf{X} = \alpha_k - \theta^*$$

$\hat{\beta}$ are maximized such that, $\{\max_{\boldsymbol{\Theta}^*} \hat{\beta} | \boldsymbol{\Theta}^* \in \text{logit}(p)\}$. Thus θ^* estimated through $\beta'\mathbf{X} \in \boldsymbol{\Theta}^*$, is a scaled estimate of θ. Let us define an arbitrary scale constant, C, such that $\theta^* = C\theta$. Then appealing to part A, it follows

$$E(\hat{\boldsymbol{\beta}}\boldsymbol{'X}) = \theta^* = C\theta = C\sum_{l=1}^{L}\sum_{g=0}^{4} w_{l,g} I(\max[X_l = g])$$

Defining $\hat{w}_{l,g}$ to be estimated through $\boldsymbol{\beta}_{\boldsymbol{l}}'X_l$, it follows that $E(\hat{\beta}'\mathbf{X}) = E(\hat{w}_{l,g})$, and thus

$$\theta^* = \sum_{l=1}^{L}\sum_{g=0}^{4} E(\hat{w}_{l,g}) I(\max[X_l = g]) = C\sum_{l=1}^{L}\sum_{g=0}^{4} w_{l,g} I(\max[X_l = g]) = C\theta$$

Then for a single toxicity type l of grade g with true weight $w_{l,g}$,

$$E(\boldsymbol{\beta}_{\boldsymbol{l}}'X_l) = E(\hat{w}_{l,g}) = Cw_{l,g} \implies E(\hat{w}_{l,g}) - Cw_{l,g} = 0 \ \forall \ l, g$$

Let a fixed position of the matrix be defined by l^* and g^*, and let $w^*_{l,g}$ and its estimate $\hat{w}^*_{l,g}$, be the weights corresponding to this position, then it follows,

$$E(\frac{\hat{w}_{l,g}}{\hat{w}^*_{l,g}}) - \frac{w_{l,g}}{w^*_{l,g}} = 0 \ \forall \ l, g$$

□

Tables

Table 1: True vs. Estimated weights for $L = 4$, $n_{(K)} = 8$, 1000 iterations

	4.1	4.2	4.3	4.4
True Weights	$\begin{bmatrix} 0.07 & 0.23 & 0.68 & 1.00 \\ 0.05 & 0.09 & 0.23 & 0.45 \\ 0.00 & 0.14 & 0.45 & 0.91 \\ 0.09 & 0.11 & 0.36 & 0.68 \end{bmatrix}$	$\begin{bmatrix} 0.05 & 0.20 & 0.55 & 0.95 \\ 0.05 & 0.25 & 0.60 & 1.00 \\ 0.13 & 0.30 & 0.50 & 0.60 \\ 0.00 & 0.10 & 0.30 & 0.70 \end{bmatrix}$	$\begin{bmatrix} 0.02 & 0.16 & 0.34 & 0.48 \\ 0.09 & 0.18 & 0.36 & 0.55 \\ 0.00 & 0.23 & 0.55 & 1.00 \\ 0.07 & 0.09 & 0.45 & 0.77 \end{bmatrix}$	$\begin{bmatrix} 0.03 & 0.16 & 0.48 & 0.56 \\ 0.00 & 0.06 & 0.42 & 1.00 \\ 0.08 & 0.12 & 0.60 & 0.80 \\ 0.12 & 0.20 & 0.24 & 0.26 \end{bmatrix}$
Cuts	(0.16, 0.31, 0.51, 0.75)	(0.18, 0.34, 0.56, 0.82)	(0.16, 0.31, 0.51, 0.75)	(0.14, 0.27, 0.45, 0.66)
Estimated Weights (Additive)	$\begin{bmatrix} 0.06 & 0.24 & 0.53 & 0.94 \\ 0.02 & 0.10 & 0.24 & 0.43 \\ 0.01 & 0.14 & 0.42 & 0.85 \\ 0.05 & 0.15 & 0.31 & 0.52 \end{bmatrix}$	$\begin{bmatrix} 0.05 & 0.20 & 0.46 & 0.82 \\ 0.06 & 0.25 & 0.55 & 0.96 \\ 0.14 & 0.30 & 0.46 & 0.64 \\ 0.00 & 0.08 & 0.28 & 0.58 \end{bmatrix}$	$\begin{bmatrix} 0.04 & 0.14 & 0.28 & 0.48 \\ 0.07 & 0.17 & 0.29 & 0.43 \\ 0.02 & 0.18 & 0.51 & 1.00 \\ 0.01 & 0.09 & 0.27 & 0.54 \end{bmatrix}$	$\begin{bmatrix} 0.06 & 0.22 & 0.49 & 0.85 \\ 0.00 & 0.10 & 0.40 & 0.87 \\ 0.03 & 0.17 & 0.42 & 0.78 \\ 0.15 & 0.24 & 0.29 & 0.29 \end{bmatrix}$
Cuts	(0.16, 0.31, 0.49, 0.72)	(0.17, 0.35, 0.56, 0.80)	(0.14, 0.27, 0.42, 0.65)	(0.18, 0.34, 0.52, 0.79)
Estimated Weights (EN Penalized)	$\begin{bmatrix} 0.01 & 0.14 & 0.48 & 0.99 \\ 0.00 & 0.03 & 0.12 & 0.29 \\ 0.00 & 0.07 & 0.32 & 0.72 \\ 0.00 & 0.05 & 0.23 & 0.52 \end{bmatrix}$	$\begin{bmatrix} 0.01 & 0.14 & 0.43 & 0.86 \\ 0.01 & 0.15 & 0.47 & 0.95 \\ 0.05 & 0.16 & 0.34 & 0.59 \\ 0.00 & 0.02 & 0.17 & 0.43 \end{bmatrix}$	$\begin{bmatrix} 0.02 & 0.09 & 0.24 & 0.46 \\ 0.03 & 0.12 & 0.30 & 0.56 \\ 0.01 & 0.13 & 0.47 & 0.98 \\ 0.00 & 0.04 & 0.31 & 0.77 \end{bmatrix}$	$\begin{bmatrix} 0.03 & 0.16 & 0.39 & 0.72 \\ 0.00 & 0.05 & 0.34 & 0.83 \\ 0.00 & 0.11 & 0.43 & 0.94 \\ 0.10 & 0.18 & 0.24 & 0.27 \end{bmatrix}$
Cuts	(0.04, 0.21, 0.44, 0.74)	(0.08, 0.25, 0.5, 0.9)	(0.06, 0.25, 0.52, 0.94)	(0.13, 0.32, 0.51, 0.94)

Table 2: Simulation results over 1000 iterations, $n_k = 8$, L=4

Additive True Weights								
Scenario	4.1		4.2		4.3		4.4	
	Subset	Full	Subset	Full	Subset	Full	Subset	Full
AIC	29		24		32		39.27	
MSPE	1.86	85.63	0.35	55.60	2.28	101.04	4.25	131.17
% Correct	95.35%	87.65%	99.12%	91.56%	94.31%	86.17%	89.38%	83.68%
% Tox Correct	99.20%	96.87%	99.79%	97.66%	99.00%	96.62%	96.59%	95.44%
% Under estimated	0.16%	2.06%	0.00%	1.63%	0.27%	2.21%	0.96%	2.38%
% Over estimated	0.64%	1.06%	0.21%	0.71%	0.77%	1.17%	2.46%	2.18%
EN Penalized True Weights								
AIC	29.92		25.88		33.26		43.88	
MSPE	5.78	99.47	2.71	96.00	4.99	119.59	5.96	135.57
% Correct	85.68%	84.36%	93.21%	85.04%	87.58%	81.71%	85.19%	79.64%
% Tox Correct	98.00%	96.12%	98.71%	95.00%	97.75%	94.42%	96.51%	95.19%
% Under estimated	0.06%	1.80%	0.06%	3.64%	0.09%	2.99%	0.73%	3.47%
% Over estimated	1.95%	2.08%	1.23%	1.36%	2.17%	2.59%	2.77%	1.34%