Web-based Supplementary Material for "Bayesian Group Sequential Clinical Trial Design using Total Toxicity Burden and Progression-Free Survival"

Brian P. Hobbs,[1,*] Peter F. Thall,[1] and Steven H. Lin[2]

[1]Department of Biostatistics, University of Texas M.D. Anderson Cancer Center, Houston, TX

[2]Department of Radiation Oncology, University of Texas M.D. Anderson Cancer Center, Houston, TX

*Corresponding Author, email: bphobbs@mdanderson.org

A Additional Discussion of Toxicity Severity Weights

The concept of TTB along with the Bayesian model and trial design provide a basis for sequential safety monitoring that accounts for both the severities and recurrences of multiple severe adverse events that may result from a therapeutic regime. The tools presented here could be valuable in settings where both safety and efficacy play critical roles, such as clinical oncology. Considering both is especially important in the context of radiation therapy, wherein recent efforts focus on developing more precise RT dosimetry algorithm and/or modalities that have the potential to limit exposure to surround host tissues but maintain effectiveness for delaying loco-regional recurrence/progression exhibited by the conventional techniques that deliver high dose to healthy tissue. Our investigation suggests that the proposed approach to safety monitoring can yield as much as a 66% increase in power and 18% reduction in mean sample size when compared to the conventional design in settings wherein toxicities derive from qualitatively diverse types of adverse events.

A.1 Elicitation Process

TTB designs require the elicitation of additional information when planning a trial. Specifically, one must assign a severity weight to each grade of each type of toxicity that is to be monitored during the course of the trial. These weights are inherently subjective. In our RT trial, they were elicited in a manner that reflects the consensus of the clinical oncologists (two radiation oncologists and one thoracic surgeon) planning and conducting the trial. A consensus was obtained using the following iterative process which required close multi-disciplinary collaboration. Our approach should be considered informal in the sense that we didn't use an established method (e.g. Hunink et al., 2014), which would have been preferable. For example, a structured communication technique known as the "Delphi method" (Dalkey and Helmer, 1963; Dalkey, 1969; Brook et al., 1986) could have been used to quantify the relative severity of each possible grade of each toxicity. Additional techniques for elicitation, characterization, and use of expert opinion can be found in Cooke (1991).

In preliminary discussions, the clinicians established the fact that comparing the relative "tolerability" of the tri-modality regimes when implemented with each RT modality was to be a primary objective of the trial (along with comparing PFS) that necessitated interim monitoring. Next we asked the clinicians to determine which types of severe adverse events (SAEs) that result from the therapeutic regime were serious enough to warrant early termination of the trial. They selected the 11 toxicities (6 radiation-induced cardiac and pulmonary SAEs and 5 postoperative complications (POCs)) given in Table 1 that are the basis for defining the TTB statistic for our trial. Additionally, at this meeting we established the evaluation periods for each toxicity. Each of the radiation-induced cardiac and pulmonary SAEs could occur as well as recur during the course of follow up as late as 52 weeks following irradiation. In contrast, the POCs were observable only at a single time-point, during the postopertaive evaluation that takes place approximately one week following surgery.

After having established the toxicities to be monitored during the trial, we met again to determine the severity levels of each toxicity. We started with severity levels based on the severity "grades" that are provided by The Common Terminology Criteria for Adverse Events (CTCAE) version 4.0 (published by The National Cancer Institute of The National Institutes of Health on May 28, 2009). CTCAE uses up to five grades to determine unique clinical descriptions of severity for adverse event reporting according to the following general guidelines: grade 1=mild without intervention, grade 2=moderate, grade 3=severe or medically significant but not immediately life-threatening, grade 4=a life-threatening condition requiring urgent intervention, and grade 5=death related to the adverse event. For example, the CTCAE provides criteria to categorize occurrences of each of the following types of toxicity at grade 1-5: PLE, RP, PNA ("aspiration"), AFIB, AL, PEM ("thromboembolic event"), and ST. Grades 2-5 are used to describe occurrences of PEF and MI, while ARDS may occur at grades 3-5.

The CTCAE criteria involve a degree of subjectivity that is unavoidable when grading an adverse event, therefore the clinicians refined the severity scales for PEF, PLE, and AL to comprise three levels defined by the manner of intervention (intervention isn't necessary for non-symptomatic toxicities of PEF and PLE). For RP, the CTCAE criteria were maintained, however the clinicians felt that it was unnecessary to distinguish between grades 1 and 2. Owing to the nature of these SAEs, the clinicians decided that all occurrences of the other toxicities following the trimodality regime should be assigned equal severity weights.

After establishing the toxicities and severity levels, we elicited the severity weights, $\boldsymbol{w}$. The weights were elicited in the range 0 to 100 since the oncologists were comfortable with this domain. However, any finite positive domain would work in practice. In our study, $w = 0$ implies no harm to the patient, while $w = 100$ represents an extent of harm that is imminently life threatening. Consensus numerical values were obtained using a two-step process. First, we obtained values from each clinician independently. Thereafter, the group met collectively to discuss the putative severity weights contributed by each clinician and select the final consensus values.

A.2 Medical Rationale for the Numerical Values used in the RT Trial

The numerical value of each severity weight reflects the relative extent of harm that is associated with experiencing the toxicity at the given level of severity in relation to the other severity levels of the same toxicity and other toxicities monitored in the trial. For example, PEF

at its highest severity level ($w = 90$) requires a major surgical procedure since it is likely that a cardiac tamponade has happened which is imminently life threatening. Symptomatic PEF requiring medical intervention necessitates aggressive diuresis, and thereby was determined to be an intermediately severe toxicity that should be assigned weight $w = 60$. Surgical intervention for PLE involves a chest tube placement, which is a more routine surgery than that of PEF. Thus, the extent of maximum harm from PLE was considered intermediate ($w = 60$) in relation to the other toxicities. PLE requiring medical intervention, which typically entails fluid restriction, was considered to be half as severe as a PEF requiring medical intervention. Perhaps the most serious postoperative complication that occurs from severing and reconnecting the esophagus is AL, a breakdown along an anastomosis causing fluids to leak. It is well established that AL leads to protracted hospital admission and can cause post-operative death. In the opinion of our clinical colleagues, an AL requiring surgical intervention was as harmful as the most severe of the other possible toxicities, and therefore assigned $w = 90$. Due to their risk of considerable and potentially permanent damage to the lungs and brain, ARDS and ST were both assigned $w = 90$. Radiation-induced damage to the pulmonary interstitium, known as RP, often requires assisted ventilation and may be irreversible at its most severe level, for which the clinicians also assigned $w = 90$. MI is potentially life threatening, and thereby a severe toxicity that can result from RT. However, owing to the fact that physicians and emergency rooms have established procedures for rescue and management of MI that avoid ICU stays and life threatening complications, it was thought to be less severe ($w = 70$) than ARDS, ST, grade 4-5 RP, and surgical interventions for AL and PEF. RI, which represents the failure of the patient to tolerate extubation following surgery, doesn't comprise a unique event according to the CTCAE, however, it is clinically important as well as potentially life-threatening, and therefore was included as a POC in our trial with $w = 70$. A PEM, or blockage to the pulmonary arteries, can be life-threatening, but prompt treatment with anticoagulant therapy can greatly reduce the risk of death. Thus, a PEM was determined to be intermediately severe ($w = 60$). PNA and AFIB are treatable, moderately severe ($w = 40$ and 30, respectively) toxicities for which we expect a relatively high incidence in the trial. While less severe, these toxicity contributed the two highest magnitudes of baseline mean toxicity burden (Table 4a) when designing the trial, and therefore are expected to play critical roles for comparing the trimodality regimes under IMRT versus PBT.

A.3 Example Hypothetical Severity Weighting Schemes

Table A mimics Table 1, with the addition of 3 sets of hypothetical severity weights for the 11 toxicities monitored in the trimodality RT trial. A higher weight corresponds to greater relative severity. In our study, for example, the event of a PLE requiring surgical intervention (w=60) was considered to be twice as undesirable as medical intervention (w=30). However, because the process is intrinsically subjective, alternative weighting schemes are conceivable. For example, HYP 1 represents a case wherein additional severity is given to conditions that involve excess fluids in the lungs and heart (PEF and PLE). The second set of hypothetical weights reflects a scenario whereby the clinicians place less emphasis on the intermediate severity levels of the three ordinal toxicities that involve medical intervention (PEF, PLE, and AL), yet maintain the relative severities of manifestations of these adverse events necessitating surgical intervention.

When combining the elicited severity weights in Table 1 with elicited incidences in Table 2, AL obtains the 3rd highest baseline expected toxicity burden (as shown in Table 4a). Thus, the

participating clinicians expect that anastomotic leak will play a critical role in determining if the trimodality regime is safer with one RT modality when compared to the other. The third hypothetical weighting set represents a scenario whereby less severity is assigned to AL. This is represented by reduced severity weights for all 3 ordinal levels when compared to their respective elicited weights.

Table A: Actual elicited and three hypothetical (HYP) severity weights for the 11 toxicities that are monitored in the esophageal cancer trial.

Recurrent Toxicities	*Severity Level*	*Elicited*	*HYP 1*	*HYP 2*	*HYP 3*
Pericardial Effusion (PEF)	non-symptomatic	10	30	10	10
	medical intervention	60	70	30	60
	surgical intervention	90	95	90	90
Pleural Effusion (PLE)	non-symptomatic	10	30	10	10
	medical intervention	30	50	15	30
	surgical intervention	60	90	60	60
Radiation Pneumonitis (RP)	grade 1-2	20	20	20	20
	grade 3	60	60	60	60
	grade 4-5	90	90	90	90
Pneumonia (PNA)	occurrence	40	40	50	40
Atrial Fibrillation (AFIB)	occurrence	30	30	40	30
Myocardial Infarction (MI)	occurrence	70	70	80	70
Postoperative Complications	*Severity Level*	*Elicited*	*HYP 1*	*HYP 2*	*HYP 3*
Anastomotic Leak (AL)	radiographic only	30	30	30	10
	medical intervention	60	60	40	30
	surgical intervention	90	90	90	70
Acute Respiratory Distress Syndrome (ARDS)	occurrence	90	90	90	90
Pulmonary Embolism (PEM)	occurrence	60	60	60	60
Reintubation (RI)	occurrence	70	70	70	70
Stroke (ST)	occurrence	90	90	90	90

B Components of the Joint Likelihood

Below we provide additional details pertaining to the probability model that is described in Section 4. Specifically, we present the four components of the likelihood contribution at the patient-level. For convenience we suppress the patient index i.

B.1 Recurrent Toxicity Processes

Given U and x, the k^{th} recurrent severity process $\{N_k(t), 0 \leq t \mid U, x\}$ is a Poisson process with conditional intensity $U\psi_k(x)$. Thus, $\mathrm{E}(N_k(t) \mid U, x) = \mathrm{var}(N_k(t) \mid U) = t\,U\,\psi_k(x)$, and $[N_{i,k}(t_i) \mid U_i, x_i]$ has Poisson pdf

$$f_N(n \mid U, x, \lambda_k, \delta_k^{\psi}) \;=\; \exp(-tU\psi_k(x))\{tU\psi_k(x)\}^n/n!, \quad n = 0, 1, \ldots.$$

Denoting the vector of baseline intensities by $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_6)$ and the vector of associated treatment effects by $\boldsymbol{\delta}^{\psi} = (\delta_1^{\psi}, \cdots, \delta_6^{\psi})$, the conditional likelihood of the multivariate counting process for recurrent toxicity $\boldsymbol{N}(t) = \{N_1(t), \cdots, N_6(t)\}$ given U is

$$\mathcal{L}_{\boldsymbol{N}}(U) = \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\delta}^{\psi} \mid \boldsymbol{N}(t), U, x) \;=\; \prod_{k=1}^{6} f_N(N_k(t) \mid U, x, \lambda_k, \delta_k^{\psi}). \tag{B.1}$$

B.2 Time-to-Surgery

Given frailty U and treatment indicator x, we assume for convenience that the time-to-surgery distribution is exponential. Assuming conditional hazard rate $\tilde{\lambda}\exp(x\tilde{\delta})/U$, the conditional likelihood is

$$\mathcal{L}_S(U) = f_S(s \mid U, x, \tilde{\lambda}, \tilde{\delta}) \;=\; \frac{\tilde{\lambda}\exp(x\tilde{\delta})}{U}\exp\{-s\tilde{\lambda}\exp(x\tilde{\delta})/U\}, \quad s > 0. \tag{B.2}$$

The probability that surgery has been performed by patient time t is

$$Pr\{S(t) > 0 \mid U, x, \tilde{\lambda}, \tilde{\delta}\} \;=\; 1 - \;\exp\{-t\tilde{\lambda}\exp(x\tilde{\delta})/U\}. \tag{B.3}$$

B.3 Severity from Postoperative Complications

The likelihood for aggregate POC severity is simply

$$\mathcal{L}_Z = \mathcal{L}\{\boldsymbol{\pi}(x) \mid \boldsymbol{Z}(t)\} \;=\; \prod_{m=1}^{24} \pi_m(x)^{Z_m(t)}. \tag{B.4}$$

B.4 Progression-free Survival

Let C denote the indicator that PFS time Y is right-censored, and let $I_g(y) = I(s_{g-1} < y \leq s_g)$. The conditional pdf and cdf of Y are

$$f_Y(y \mid \boldsymbol{\gamma}, \delta^{\xi}, U, x) = \prod_{g=1}^{G} \xi_g(x, U)^{I_b(y)} \exp\left[-I_g(y)\left\{\xi_g(x, U)\,(y - s_{g-1}) + \sum_{l=1}^{g-1} \xi_l(x, U)(s_l - s_{l-1})\right\}\right], \tag{B.5}$$

and

$$F_Y(y|\boldsymbol{\gamma}, \delta^{\xi}, U, x) = 1 - \exp\left[-\sum_{g=1}^{G} I_g(y)\left\{\xi_g(x, U)\,(y - s_{g-1}) + \sum_{l=1}^{g-1} \xi_l(x, U)(s_l - s_{l-1})\right\}\right]. \tag{B.6}$$

Thus, each patient's PFS likelihood contribution is

$$\mathcal{L}_Y(U) = \mathcal{L}(\boldsymbol{\gamma}, \delta^{\xi} \mid Y, C, U, x) = f_Y(Y, C \mid \boldsymbol{\gamma}, \delta^{\xi}, U, x)^{1-C}\{1 - F(Y, C \mid \boldsymbol{\gamma}, \delta^{\xi}, U, x)\}^{C}. \tag{B.7}$$

C Bayesian Computation

MCMC was implemented using both Gibbs sampling and Metropolis−Hastings algorithms. A total of 30,000 updates sampled from each of four parallel chains following burn-in periods of 30,000 was sufficient to obtain convergence, following Gelman and Rubin (1992), for all simulated scenarios. Below we provide full conditional distributions, denoted by $q(\cdot)$, as functions of model parameters $\boldsymbol{\theta}$, frailties $\boldsymbol{U}$, patient time $\boldsymbol{t}$, treatment assignments $\boldsymbol{x}$, observables $\boldsymbol{\mathcal{D}}$, and fixed hyperparameters, as discussed in Section 4.5. The following explains how we sampled from the posterior distribution for $\Delta(\boldsymbol{\theta}, \boldsymbol{w})$ from post-processing of the MCMC draws of $\boldsymbol{\theta}|\mathcal{D}$.

C.1 Baseline Parameters

Gibbs sampling is possible for posterior estimation of baseline event rates for toxicity recurrence and surgery, which were assumed to have conditionally conjugate $\Gamma-$prior distributions. We denote the pair of hyperparameters for the prior mean and rate by (l_r, c^ψ) for recurrent toxicity severity level r and $(\tilde{l}, \tilde{c})$ for surgery. Denoting $N_{i,r} = N_{i,r}(t_i)$, $S_i = S_i(t_i)$, and $Z_{i,m} = Z_{i,m}(t_i)$, the resulting full conditional posteriors for λ_r and $\tilde{\lambda}$ are proportional to

$$q(\lambda_r \mid \delta_r^\psi, \boldsymbol{U}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\mathcal{D}}, l_r, c^\psi) \propto \Gamma\left(l_r c^\psi + \sum_{i=1}^n N_{i,r}\ ,\ c^\psi + \sum_{i=1}^n t_i U_i \exp(-x_i \delta_r^\psi)\right), \tag{C.1}$$

and

$$q(\tilde{\lambda} \mid \tilde{\delta}, \boldsymbol{U}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\mathcal{D}}, \tilde{l}, \tilde{c}) \propto \Gamma\left(\tilde{l}\tilde{c} + \sum_{i=1}^n S_i\ ,\ \tilde{c} + \sum_{i=1}^n t_i \exp(-x_i\tilde{\delta})/U_i\right). \tag{C.2}$$

For the PFS hazard, define $\xi_{i,g} = \xi_g(x_i, U_i) = U_i \gamma_g \exp(-x_i \delta^\xi)$, and $\boldsymbol{\xi}_i = (\xi_{i,1}, \cdots, \xi_{i,G})$. Under the piecewise exponential model, the hazard at patient time t is piecewise constant, $h(t, \boldsymbol{\xi}_i) = \xi_{i,g}$, for $s_{g-1} < t \leq s_g$, and the cumulative hazard at patient time t is

$$H(t, \boldsymbol{\xi}_i) = \sum_{g=1}^G [\{\ \min(\ \max(t, s_{g-1}), s_g) - s_{g-1}\ \}\ \xi_{i,g}].$$

We can express the likelihood function for PFS as $\prod_{i=1}^n h(t, \boldsymbol{\xi}_i)^{1-C_i} \exp\{-H(t, \boldsymbol{\xi}_i)\}$. Let c^γ denote the Γ-prior rate hyperparameter, and κ_g denote the Γ-prior mean for baseline hazard in gth interval, where $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_G)$. Metropolis−Hastings was used to estimate the posterior of each piecewise constant hazard. Recalling that $I_g(t) = I(s_{g-1} < t \leq s_g)$, the full conditional posterior for γ_g is

$$q(\gamma_g \mid \delta^\xi, \boldsymbol{U}, \boldsymbol{x}, \boldsymbol{\mathcal{D}}, c^\gamma, \boldsymbol{\kappa}\ ) \propto \exp\{-H(Y_i, \boldsymbol{\xi}_i)\}\ \Gamma\left(\gamma_g \mid \kappa_g c^\gamma + \sum_{i=1}^n I_g(Y_i)(1 - C_i)\ ,\ c^\gamma\right). \tag{C.3}$$

Let $I_x(x_i)$ indicate that $x_i = x$. Given the set of Dirichlet concentration hyperparameters, $\boldsymbol{p} = (p_1, \cdots, p_{24})$, the posterior for the probability of POC severity for modality x is

$$q\{\pi(x) \mid \boldsymbol{x}, \boldsymbol{\mathcal{D}}, \boldsymbol{p}\} = Dirichlet\left(\ p_1 + \sum_{i=1}^n I_x(x_i) Z_{i,1}, \cdots, p_{24} + \sum_{i=1}^n I_x(x_i) Z_{i,24}\ \right). \tag{C.4}$$

C.2 Modality Effects

Because our model precludes a conditionally conjugate posterior for $\boldsymbol{\delta}$, we did posterior estimation by rejection sampling. Let $\tilde{v}$ and v^{ξ} denote prior variance hyperparameters for $\tilde{\delta}$ and δ^{ξ}, respectively. Metropolis−Hastings was implemented using the following full conditional posteriors for $\tilde{\delta}$, δ^{ξ}, and δ_r:

$$q(\tilde{\delta} \mid \tilde{\lambda}, \boldsymbol{U}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\mathcal{D}}, \tilde{v}) \propto \exp\left(-\tilde{\lambda}\sum_{i=1}^{n} t_i \exp(-x_i\tilde{\delta})/U_i\right)\ N\left(\tilde{\delta} \mid 2\tilde{v}\sum_{i=1}^{n} x_i S_i\ ,\ \tilde{v}\right), \tag{C.5}$$

$$q(\delta^{\xi} \mid \boldsymbol{\gamma}, \boldsymbol{U}, \boldsymbol{x}, \boldsymbol{\mathcal{D}}, v^{\xi}\ ) \propto \exp\left\{-H(t, \boldsymbol{\xi}_i)\right\}\ N\left(\delta^{\xi} \mid 2v^{\xi}\sum_{i=1}^{n} x_i I_g(Y_i)(1-C_i)\ ,\ v^{\xi}\right), \tag{C.6}$$

and

$$q(\delta_r^{\psi} \mid \lambda_r, \boldsymbol{U}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\mathcal{D}}, \delta^{\psi}, \omega) \propto \exp\left(-\lambda_r\sum_{i=1}^{n} t_i U_i \exp(-x_i\delta_r^{\psi})\right)\ N\left(\delta_r^{\psi} \mid \delta^{\psi} - \omega^2\sum_{i=1}^{n} x_i N_{i,r}\ ,\ \omega^2\right). \tag{C.7}$$

Let m^{ψ} denote the number of unique severities and v^{ψ} denote the hyperprior variance for δ^{ψ}. The full conditional posterior for the hierarchical mean δ^{ψ} is normal with mean the average of severity-specific modality effects

$$q(\delta^{\psi} \mid \lambda_r, \boldsymbol{\delta}^{\psi}, \omega, m^{\psi}, v^{\psi}) \propto N\left\{\delta^{\psi} \mid \frac{\frac{1}{\omega^2}\sum_{r=1}^{m^{\psi}}\delta_r^{\psi}}{\frac{m^{\psi}}{\omega^2}+\frac{1}{v^{\psi}}}\ ,\ \left(\frac{m^{\psi}}{\omega^2}+\frac{1}{v^{\psi}}\right)^{-1}\right\}. \tag{C.8}$$

Let $\mathbf{1}_m$ denote a vector of 1s of length m. The full conditional distribution for the hierarchical variance ω^2 is proportion to

$$q(\omega \mid \delta^{\psi}, \boldsymbol{\delta}^{\psi}) \propto I(\omega > 0)I(\omega \leq 10)(1/\omega)\exp\left\{-\frac{1}{2\omega^2}(\boldsymbol{\delta}^{\psi} - \delta^{\psi}\mathbf{1}_{m^{\psi}})'(\boldsymbol{\delta}^{\psi} - \delta^{\psi}\mathbf{1}_{m^{\psi}})\right\} \tag{C.9}$$

C.3 Frailties and Frailty Variance

Define $A_i(\phi) = \sum_{r=1}^{m^{\psi}} N_{i,r} - (S_i + C_i + 1/\phi)$. Given the data and frailty variance, ϕ, the full conditional distribution for predicting the ith patient's latent frailty U_i is proportional to

$$q(U_i \mid \boldsymbol{\lambda}, \tilde{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\delta}^{\psi}, \tilde{\delta}, \delta^{\xi}, \phi, \boldsymbol{\mathcal{D}}) \propto \tag{C.10}$$
$$U_i^{A_i(\phi)}\exp\left\{-t_i\left(U_i\sum_{r=1}^{m^{\psi}}\lambda_r\exp(-x_i\delta_r^{\psi})\ +\ \frac{1}{U_i}\tilde{\lambda}\exp(-x_i\tilde{\delta})\right)\ -\ H(Y_i, \boldsymbol{\xi}_i)\ -\ \frac{1/\phi+1}{U_i}\right\}.$$

The full conditional distribution for the frailty variance given the set of all patient frailties, $\boldsymbol{U}$, is proportional to

$$q(\phi \mid \boldsymbol{U}) \propto I(\phi > 0)I(\phi \leq 10)\prod_{i=1}^{n}\Gamma^{-1}(U_i \mid \frac{1}{\phi}+2\ ,\ \frac{1}{\phi}+1). \tag{C.11}$$

C.4 Mean TTB Difference

Recall that $x = -0.5$ for IMRT and $x = 0.5$ for PBT. Following from (5), (6), and (7), the J^{th} sample of $\Delta(\boldsymbol{\theta}, \boldsymbol{w})$ is obtained by mapping the J^{th} sample of $\boldsymbol{\theta}$ into

$$\begin{aligned}\Delta(\boldsymbol{\theta}, \boldsymbol{w})^{(J)} = &\sum_{r=1}^{m^{\psi}} w_r^* \lambda_r^{(J)} \left\{\exp(0.5\delta_r^{\psi(J)}) - \exp(-0.5\delta_r^{\psi(J)})\right\} \\ &+ \tilde{\boldsymbol{w}}'\boldsymbol{\pi}(0.5)^{(J)} \left\{1 - \left(\frac{\phi^{(J)}\tilde{\lambda}^{(J)}\exp(0.5\tilde{\delta}^{(J)})}{\phi^{(J)}+1} + 1\right)^{-(1/\phi^{(J)}+2)}\right\} \\ &- \tilde{\boldsymbol{w}}'\boldsymbol{\pi}(-0.5)^{(J)} \left\{1 - \left(\frac{\phi^{(J)}\tilde{\lambda}^{(J)}\exp(-0.5\tilde{\delta}^{(J)})}{\phi^{(J)}+1} + 1\right)^{-(1/\phi^{(J)}+2)}\right\}.\end{aligned} \tag{C.12}$$

References

Brook, R. H., Chassin, M. R., Fink, A., Solomon, D. H., Kosecoff, J., and Park, R. E. (1986), "A method for the detailed assessment of the appropriateness of medical technologies," *International Journal of Technology Assessment and Health Care*, 2, 53–63.

Cooke, R. M. (1991), *Experts in Uncertainty: Opinion and Subjective Probability in Science. Environmental Ethics and Science Policy Series*, New York: Oxford University Press.

Dalkey, N. and Helmer, O. (1963), "An Experimental Application of the Delphi Method to the use of experts," *Management Science*, 9, 458467.

Dalkey, N. C. (1969), "An experimental study of group opinion," *Futures*, 1, 408426.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences (with discussion)," *Statistical Science*, 7, 457–511.

Hunink, M. G. M., Weinstein, M. C., Wittenberg, E., Pliskin, J. S., Drummond, M. F., Glasziou, P. P., and Wong, J. B. (2014), *Decision Making in Health and Medicine: Integrating Evidence and Values*, Cambridge: Cambridge University Press.