Received ???; | Revised ???; | Accepted ???

DOI: xxx/xxxx

RESEARCH ARTICLE

Bayesian variable selection based on clinical relevance weights in small sample studies - Application to colon cancer

Sandrine Boulet*[1] | Moreno Ursino[1] | Peter Thall[2] | Anne-Sophie Jannot[1,3] | Sarah Zohar[1]

[1]INSERM U1138, Team 22, Centre de Recherche des Cordeliers, University Paris Descartes, University Pierre et Marie Curie, Paris, France

[2]Department of Biostatistics, M.D. Anderson Cancer Center, Houston, USA

[3]Department of Statistics, Computing and Public Health, Hôpital européen Georges-Pompidou, AP-HP, Paris, France

Correspondence

*Sandrine Boulet, Email: sandrine.boulet@crc.jussieu.fr

Summary

Using clinical data to model the medical decisions behind sequential treatment actions raises methodological challenges. Physicians often have access to many covariates that may be used when making sequential treatment decisions for individual patients. Statistical variable selection methods may help finding which of these variables are used for this decision in everyday practice. When the sample size is not large, Bayesian variable selection methods can address this setting and allow for expert information to be incorporated into prior distributions. Motivated by clinical practice data involving repeated dose adaptation for Irinotecan in colorectal metastatic cancer, we propose a modification of the Stochastic Search Variable Selection (SSVS) method, which we call Weight-based SSVS (WBS). We use clinical relevance weights elicited from physician experts to construct prior distributions, with the goal to identify the most influential toxicities and other covariates used for dose adjustment. We evaluate and compare the WBS model performance to the Lasso and SSVS through an extensive simulation study. The simulations show that WBS has better performance and lower rates of false positives and false negatives than the other methods but depends strongly on the covariate weights.

KEYWORDS:

Stochastic search variable selection, Clinical relevance weights elicitation, Informative priors, Repeated measures.

1 | INTRODUCTION

Recent innovations in the accessibility and availability of hospital electronic health records (EHRs), where comprehensive information about patients are gathered, have made this source of structured data accessible for statistical analysis and medical research[1,2]. Reuse of such records facilitates a better understanding of medical decision making associated with sequential treatment decisions. Modeling the medical decisions behind treatment actions as a function of patient characteristics, including previous treatments or doses and clinical outcomes, raises methodological challenges. Statistical models must account for many possible covariates, comorbidities, and clinical events that may explain dose modifications or sequences of treatments for individual patients.

General treatment recommendations often are based on clinical trial results, in which patients are a selected subset of the actual treated population. Inclusion and exclusion criteria of clinical trials can be so strict that older, pediatric, or patients with particular comorbidities may be excluded. As a result, general recommendations may be unsuitable for these patients, which

[0]**Abbreviations:** SSVS, stochastic search variable selection; RMSE, root mean square error; FPR, false positive rate; FNR, false negative rate

forces practicing physicians to modify treatment decisions based on each patient's characteristics and history. In medical practice, at each patient visit, the physician must choose a new treatment or modify the dose of an ongoing treatment based on their experience, outside recommendations, and the patient's current history, including covariates, previous treatments, and clinical outcomes. Modeling such medical decision making based on observed toxicities and covariates can be difficult, particularly because of the multidimensionality of clinical toxicity observations, classified by organ and severity (https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm). Physicians typically prioritize certain covariates over others and may use their own subjective clinical relevance weights when making treatment decisions. Accounting for clinical relevance weights may be highly informative when analyzing EHR data to estimate optimal treatment sequences or dose modifications.

In oncology, many innovative molecules and biomarkers (genomic, clinical, or physiological) have been recently proposed, which has led to consideration of many small subpopulations of diseases previously assumed to be homogeneous[3]. In such small subpopulations, identifying causes of toxicities and resulting treatment modifications in everyday health care may be quite complicated. With small samples, fitted multivariate models may have large variances of estimated parameters, limiting inferential reliability. Variable selection methods in such settings may mitigate this difficulty by keeping only covariates having the strongest effects. Most of these methods are frequentist, using penalization criteria, such as the bridge, Lasso, SCAD, LARS, elastic net, or OSCAR regressions[4,5,6,7,8]. One of the most popular methods is the Lasso, which performs variable selection by applying an L1-penalty to least squares. The Lasso shrinks some coefficients to zero, while selecting others. However, these variable selection methods have been validated and associated with good convergence properties for large samples. In our setting, given the small sample sizes and large number of variables considered, these methods often fail to provide reliable results.

Bayesian regression is a reasonable alternative approach when working with small samples. In the Bayesian framework, variable selection may be performed by applying so-called "spike and slab" priors to identify and estimate the posteriors of nonzero regression parameters[9], while deleting variables that do not appear to predict outcome. Mitchell and Beauchamp assumed the prior distribution of each regression coefficient to be a two-component mixture of a point mass at 0 (the spike) and a uniform distribution on a finite interval (the slab)[10]. These two components quantify prior belief about whether a covariate effect exists and, if it does, the effect's magnitude. This method has been extended in a variety of ways, focusing on specific formulations of priors[11,12,13,14,15]. A very useful extension is the stochastic search variable selection (SSVS) method, in which the prior of each parameter is assumed to be a mixture of two normal distributions centered at zero but with very different variances[11], one large and the other small. Other methods specify continuous prior distributions that approximate the "spike and slab" shape while shrinking the estimates toward zero[16,17]. Park and Casella extended the frequentist Lasso by assuming a double exponential distribution for the regression parameters[17]. For a review of Bayesian variable selection methods, see O'Hara and Sillanpää[9].

A major advantage of Bayesian inference is that external information can be incorporated formally into prior distributions. Many Bayesian variable selection studies use priors that do not specify a preference for any variable but may, for instance, show a preference for sparse models. Several Bayesian variable selection studies have proposed simple approaches for incorporating prior knowledge, by independently assigning each variable a prior probability of being included in the model[18]. For example, Kitchen et al. provide a method for prior parameter specification using the scientific literature for regression problems with many predictors in which the priors include both shrinkage and variable selection components, extending the Kuo and Mallick model[12]. The authors posit this approach as a solution to the phenotype-genotype problem. They construct a so-called "voting" prior by assigning a count to each codon position based on the number of times that it is identified as important in the literature[19]. In such a study, the same amount of information is available for each biomarker. However, in other settings with many covariates, such as clinical variables, the variable selection literature is limited.

We are motivated by clinical practice data in which dose modifications were made over multiple cycles for patients being treated with Irinotecan for colorectal cancer. We use elicited clinical relevance weights to reflect the medical experience of practicing physicians[20]. When a physician makes a decision regarding a patient's treatment or dose modification, certain patient characteristics and previous clinical outcomes are given more weight than others. This practice implies that the decision is influenced by the physician's personal experience and knowledge regarding the clinical relevance of certain covariates. To analyze our motivating data set, we elicited expert relevance weights for patient covariates, including toxicities, from each of several physicians and used these weights to construct prior distributions.

In colorectal cancer setting, biomarkers that explain certain toxicities have recently been identified. For example, activating mutations of the KRAS oncogene have been shown to be predictive of resistance of anti-EGFR agents, such as Cetuximab, which is sometimes used with Irinotecan in treating metastatic colorectal cancer[21]. Another example is the gene UGT1A1, which is strongly associated with bilirubin levels, for which a variant is linked to an increased risk of Irinotecan toxicity. This finding suggests that patient subgroups defined by these variables should be treated differently. The most common adverse events associated

with Irinotecan include vomiting, nausea, diarrhea, asthenia, neutropenia and anemia. All of these toxicities occur at various severity levels, classified as grade 0, 1, 2, 3, or 4. Generally, the dose for each patient is calculated according to his/her body surface area. Toxicities often require decreasing the patient's dose, extending the delay between two dose administration intervals (or cycles), or temporarily discontinuing treatment. In addition, age, performance status, bilirubin, genetic polymorphism, and drug administration schedule have been shown to be related to Irinotecan toxicity[22]. While standard protocols for dose adjustment in chemotherapy for adverse events exist, these recommendations do not account for possible associations between the covariates noted above. Thus, in clinical practice, physicians adapt doses and schedules during successive cycles of treatment by accounting for all of the patient's characteristics and treatment history, including doses and outcomes. As a result, each patient's dose regimen over multiple cycles often differs from what is recommended in standard protocols.

To analyze the Irinotecan data, we proceeded in four steps : (1) First, we asked each of four physicians who are experts in treating colorectal cancer to specify numerical clinical relevance weights to reflect their beliefs about the importance of each variable in their decision making. When a covariate takes on a specific value, physicians will reduce the dose with a probability that depends on the physician's clinical relevance weight associated with this covariate. Each physician provided their clinical relevance weight associated with each grade of each toxicity type and each level of each covariate. (2) Next, we constructed a modification of the SSVS method, which we call weight-based SSVS (WBS), to identify covariates that influence dose adjustment. We used the elicited clinical relevance weights to compute the hyper-parameters of the prior distributions of the variables' inclusion indicators to provide a basis for the WBS. (3) Third, we compared the influence of the physicians' clinical relevance weights using our proposed method with two other methods, the Lasso and conventional SSVS. We also studied the effect of sample size on our method. (4) Finally, we applied our method to analyze the Irinotecan data set.

The remainder of the paper is organized as follows. Methods and criteria for model comparison are discussed in section 2. In section 3, we describe a simulation study design to assess our methodology, and we present the simulation results in section 4. We present the data analysis of our case study in section 5 and close with a discussion in section 6.

2 | METHODS

Our motivating data set is made of one continuous dependent variable, the dose of Irinotecan (mg/m^2), and twelve covariates: age > 80 years, weight loss > 10% since the beginning of treatment, WHO (World Health Organization) score, bilirubin > 35 μmol/L, treatment line $\geq$ 3 (that is, the patient previously received more than 3 other cancer treatments) and toxicities associated with Irinotecan treatment (vomiting, nausea, diarrhea, asthenia, neutropenia, thrombopenia and anemia). All these variables are known for each patient and each of his/her cycles, so it is repeated data. In cancer, a cycle is a course of treatment repeated on a regular schedule with periods of rest in between. In our setting, the treatment is given on day 1 and the next cycle begins 14 days after. Thus, a new cycle starts when the patient receives a new dose. Side effects can appear during the cycle, as described in Figure 1 . The data set was built to study the impact of covariates on clinicians' decision to adapt the doses, therefore each line of the dataset corresponds to one cycle of a given patient and contains the dose given to the patient at the beginning of the cycle and the observed covariates between the start of the previous cycle and the present cycle.

The following notation is introduced for the general problem of variable selection using clinical relevance weights for longitudinal data. Let n be the number of subjects, J the number of covariates, and K the number of cycles. For patient $i \in \{1, ..., n\}$, let $\mathbf{x}_{i,k}$ be the J-vector of covariates available at the start of cycle $k \in \{1, ..., K\}$. In our application, certain variables, such as age and treatment, are assumed to be fixed in time, while other patient characteristics and toxicities associated with adverse events and treatment responses may change over time. For $j \in \{1, ..., J\}$, let $x_{i,k,j}$ take values in $\{x'_{j,0}, x'_{j,1}, ..., x'_{j,C_j}\}$, where $x'_{j,c}$ is the cth most severe level, $c = 0, ..., C_j$. For instance, age is dichotomized and takes on values in $\{0, 1\}$, where 0 means that age < 80, and 1 means that age $\geq$ 80. Each toxicity is categorized by integer-valued grades between 0 (absent) and either 3 or 4 (most severe). Let $\mathbf{z}_{i,k}$ be the vector of all patient covariates in $\mathbf{x}_{i,k}$, using the dummy coding of categorical and/or ordinal data, that is, $\mathbf{z}_{i,k}$ has length $L = \sum_{j=1}^{J} C_j$ with each $x_{i,k,j}$ split into C_j dummies variables, each taking a value of 1 if the corresponding category was selected in $x_{i,k,j}$ and 0 otherwise. In the following section, we formally define clinical relevance weights and explain how we use them to determine prior distribution hyper-parameters in a Bayesian model.

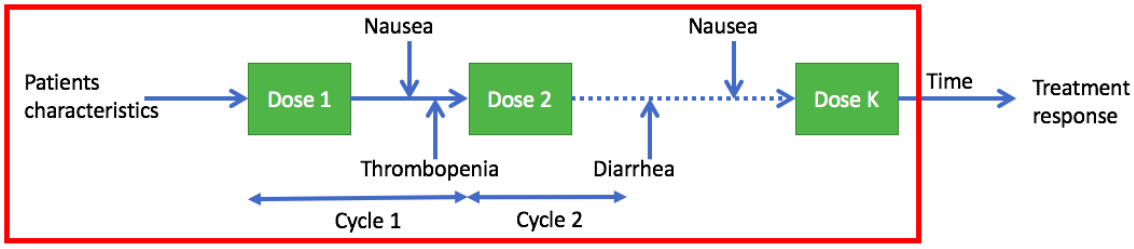


FIGURE 1 Example for one patient of dose adjustment according to patients' characteristics and toxicities

2.1 | Clinical relevance weights

Clinical relevant weights reflect the importance that a physician places on each value of each variable used in a medical decision. Once these weights are elicited, they can be used to compute the hyper-parameters of the prior distributions of each variable's inclusion indicator in the Bayesian model. In the elicitation, each of several physicians who are experts in the field are asked to provide positive-valued numerical weights $\mathbf{w}_j$ for each variable $j \in \{1, ..., J\}$ and each of its possible values $c \in \{1, ..., C_j\}$. Denote $\mathbf{w}_j = (w_{j,1}, ..., w_{j,C_j})$, ordered such that $0 = w_{j,0} \leq w_{j,1} \leq ... \leq w_{j,C_j} = W_{max}$ for each j. The value $w_{j,c}$ reflects how much the expert thinks he/she would change the dose on a scale between 0 and W_{max} when the severity level $x_{j,c}$ is observed. The value $w_{j,0}$ is not included in the weight vector because it represents the reference level, that is, no toxicities in the case of toxicity variables or the reference value in the case of a dichotomous variable. For example, letting $W_{max} = 100$, the vector $w_6 = \{0, 20, 80, 90\}$ refers to clinical relevance weights for the 6th variable, which has $C_6 = 4$ levels. By analogy with $\mathbf{z}$, we define $\mathbf{w}'$ to be the L-vector of all clinical relevance weights. This structure follows that structure introduced by Bekele and Thall,[23], who used elicited toxicity severity weights to define total toxicity burden for use in phase I dose-finding.

For our setting, oncologists helped us build a questionnaire by choosing the variables and thresholds that they used for decision making. Different experts could give different weights, which led to different prior distributions. Toxicity covariates were defined using the Common Terminology Criteria for Adverse Events (CTCAE)'s grades. Four clinicians separately completed the online questionnaire. Each clinician specified a numerical clinical relevance weight for each grade of each toxicity type and each level of each covariate, within the range $[0, W_{max}] = [0, 100]$, with 0 corresponding to no severity and 100 the highest possible severity.

Table 1 presents the elicited clinical relevance weights for each oncologist and each variable. For example, $\{0, 20, 80, 90\}$ in the row "vomiting" of Table 1 refers to clinical relevance weights of grades $\{1, 2, 3, 4\}$ given by clinician 1. The table suggests that if grade 2 vomiting is observed clinician 1 reduces the dose in 20 % of cases. For all clinicians except clinician 1, the WHO score is very often considered for dose adjustment, and grade 4 is always considered for dose reduction. Clinician 1 gives the small weight of 20 to WHO scores 2, 3, and 4. Clinicians 2, 3 and 4 do not consider treatment line for decision making, while clinician 1 considers it for dose reduction. Concerning toxicities, vomiting is used very differently by the four clinicians to make decisions. Finally, on average, clinicians 2 and 3 appear to assign lower weights to many levels of many variables (cf. sum of weights in Table 1) compared with clinicians 1 and 4. The elicited weights suggest that the four physicians are likely to act differently when facing the same situations.

2.2 | Weight-based SSVS (WBS)

The general framework of clinical practice reflected by the Irinotecan data is repeated dose adaptation over successive treatment cycles based on patients' most recently updated treatment, dose, and outcome data. In the first cycle, a starting dose is chosen based on the patients' baseline characteristics, including age, bilirubin, WHO score, and possibly previous treatments. Thereafter, if toxicities are observed in a given cycle, the clinician may adapt the dose for the following cycle. We assume the following linear mixed effect model, which is intended to use the updated covariates in each cycle to explain the physician's decisions. Let $d_{i,k}$ denote the dose given to patient i at cycle k and $\mathbf{z}_{i,k}$ the vector of the patient's covariates at that cycle visit for $i \in \{1, ..., n\}$ and $k \in \{1, ..., K\}$. The model is given by the equation

$$d_{i,k} = \theta_0 + \boldsymbol{\theta}^T \mathbf{z}_{i,k} + \gamma_i + \epsilon_{i,k} \tag{1}$$

TABLE 1 Clinical relevance weights for each variable elicited from each clinician.

	Clinicians															
Covariates	**1**				**2**				**3**				**4**			
Age ≥ 80 years	100	-	-	-	60	-	-	-	100	-	-	-	80	-	-	-
Treatment line 3, > 3	30	50	-	-	0	0	-	-	0	0	-	-	0	0	-	-
Weight loss > 10%	50	-	-	-	20	-	-	-	50	-	-	-	80	-	-	-
WHO score (1,2,3,4)	0	20	20	20	0	0	40	100	0	0	80	100	0	20	80	100
Bilirubin >35 μmol/L	100	-	-	-	40	-	-	-	100	-	-	-	20	-	-	-
Toxicity grades 1, 2, 3, 4																
Vomiting	0	20	80	90	0	30	70	100	0	10	10	10	10	20	80	100
Nausea	0	20	80	-	0	10	50	-	0	10	10	-	10	30	80	-
Diarrhea	0	40	80	100	0	20	50	100	0	50	80	100	0	20	70	90
Asthenia	10	50	100	-	10	10	40	-	0	0	70	-	10	50	70	-
Neutropenia	0	70	100	100	0	0	30	50	0	0	50	50	0	20	70	80
Thrombopenia	40	100	100	100	0	0	20	30	0	0	50	50	0	50	80	100
Anemia	0	50	80	100	0	0	20	30	0	0	0	0	0	20	50	70
Sum of weights	1900				930				980				1560			

where $\boldsymbol{\theta}$ is a parameter vector of length L; $\gamma_1, \cdots, \gamma_n$ are iid $\mathcal{N}(0, \sigma_\gamma^2)$ random patient effects; and $\epsilon_{i,k} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the error term. In this formulation, the Markov property is assumed conditional on the random effects such that the dose given at cycle k depends only on the covariates resulting from the previous cycle and collected in $\mathbf{z}_{i,k}$. Thus, $\mathbf{z}_{i,k}$ includes indicators of the grades of any toxicities that occurred in cycle $k-1$ and any updated values of the baseline covariates. Note that toxicities are not the outcomes of the regression, but rather they are the covariates.

Our objective is to find a subset, $\mathcal{I}$, of the vector $\mathbf{z}$ that describes the data well. For covariate index $l \in \{1, ..., L\}$, let I_l to be the indicator variable that takes the values

$$I_l = \begin{cases} 1 & \text{if } z_{..l} \text{ is in the model} \\ 0 & \text{otherwise,} \end{cases}$$

where we refer to the l^{th} covariate by $z_{..l}$ for convenience. The SSVS approach assumes that the prior distribution for each covariate parameter in the model (1) is a mixture of two normal distributions, both centered at zero but with different variances. Specifically, as given in [11],

$$\theta_l | I_l \sim (1 - I_l)\mathcal{N}(0, \tau_l^2) + I_l \mathcal{N}(0, g_l \tau_l^2) \tag{2}$$

where $\mathbb{P}(I_l = 1) = p_l$, denoted by $I_l \sim \mathcal{B}(p_l)$. In equation eq. (2), the first term (the spike) accounts for the non-selected covariate $z_{..l}$ ($I_l = 0$), with the density centered at 0 having small variance. Note that the sequence of zero-centered normal distributions with variance zero as limit is the Dirac delta. Thus, rather to use a prior with a point mass on the regression coefficient equaling 0, we decided to use a normal distribution with small variance to approximate it and thus allow us to use usual statistical software. The second term accounts for the selected covariates $z_{..l}$, where $I_l = 1$, and therefore, it has a large variance.

In the classical SSVS framework, p_l is set to 0.5. Here, we use the elicited weights to choose a prior distribution for I_l such way that I_l is more likely to be 1 when its elicited clinical relevance weight is high, and it is more likely to be 0 otherwise. Rather than fixing p_l, we assume that it follows a beta prior distribution,

$$p_l \sim Beta(a_l, b_l). \tag{3}$$

Consequently, $I_l | p_l$ follows a beta-binomiale distribution of parameters (a_l,b_l). The main innovation of our proposed Weight-based Stochastic search variable selection (WBS) method is that we assume the prior equation given by eq. (3) and compute its hyper-parameters a_l and b_l of using the elicited clinical relevance weights w_l' described in subsection 2.1. Thus, first, each clinician separately specifies a vector of numerical clinical relevance weights $\mathbf{w}'$ associated with each grade of each toxicity type and each level of each covariate. Then, we find a_l and b_l by solving the equations

$$\begin{cases} \mathbb{E}(p_l) = \dfrac{a_l}{a_l + b_l} = \dfrac{w_l'}{W_{max}} \\ \mathbb{V}(p_l) = \dfrac{a_l b_l}{(a_l + b_l)^2 (a_l + b_l + 1)} = S. \end{cases}$$

TABLE 2 (a, b) values giving a beta(a, b) distribution centered at clinical relevance weight/100

Clinical relevance weights	0	10	20	30	40	50	60	70	80	90	100
a	1.41	1.41	1.41	2.88	4.43	5.78	6.65	6.72	5.64	5.64	5.64
b	5.64	5.64	5.64	6.72	6.65	5.78	4.43	2.88	1.41	1.41	1.41

Thus, the mean of p_l is w_l'/W_{max}, while S is a tuning parameter that accounts for uncertainty, chosen based on a preliminary sensitivity analysis. In our application, we choose the same S for all prior distributions, although in general the value may be different for different p_l, reflecting different levels of the physician's confidence for different covariates. Note that no inference procedures are needed for the clinical relevance weights elicited from each clinician. To implement the WBS method, we set the minimal and maximal thresholds to 20 and 80, respectively, to avoid highly informative priors. Values of a and b obtained from clinical relevance weights used in the WBS method are presented in Table 3 and Figure 2 . We choose $W_{max} = 100$ because the scale $[0, 100]$ is well known and easily understandable by clinicians, and we set $S = 0.02$ based on our preliminary sensitivity analysis. S quantifies the variability that one wants to add to the prior distribution. We denote the WBS method based on the m^{th} clinician's clinical relevance weights as WBS(m), for $m \in \{1, 2, 3, 4\}$.

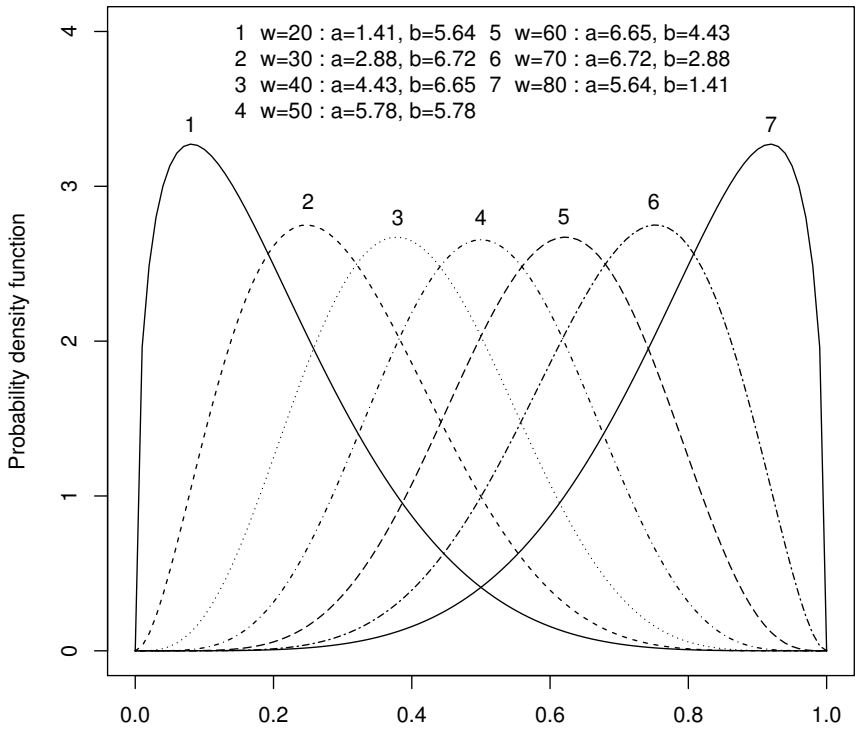


FIGURE 2 Probability density functions for p's beta distributions with parameters a and b and mean w/100

2.3 | Combination of prior opinions

To combine prior opinions from different physicians, we propose two methods: (1) use of the WBS method with a mixture prior obtained by weighting each of the four elicited physician priors 0.25 each, and (2) Bayesian Model Averaging (BMA) of the WBS models used with each clinician's clinical relevance weights. To accomodate inconsistent weighting by different physicians, BMA calculates the predictive distribution of a coefficient $\theta_l, \in \{1, ..., L\}$, as a weighted average of posterior distributions of θ_l using each model WBS(m), $m \in \{1, ..., 4\}$[24]:

$$p(\theta_l|\mathbf{d}, \mathbf{z}) = \sum_{m=1}^{4} p(WBS(m)|\mathbf{d}, \mathbf{z}) p(\theta_l|WBS(m), \mathbf{d}, \mathbf{z})$$

where the weights $p(WBS(m)|\mathbf{d}, \mathbf{z})$ are the posterior probabilities of each model:

$$p(WBS(m)|\mathbf{d}, \mathbf{z}) = \frac{p(\mathbf{d}, \mathbf{z}|WBS(m))p(WBS(m))}{\sum_{m_1=1}^{4} p(\mathbf{d}, \mathbf{z}|WBS(m_1))p(WBS(m_1))}$$

The marginal likelihood of the model $WBS(m)$ is:

$$p(\mathbf{d}, \mathbf{z}|WBS(m)) = \int p(\mathbf{d}, \mathbf{z}|\theta, WBS(m))p(\theta|WBS(m))d\theta$$

where $p(\mathbf{d}, \mathbf{z}|\theta, WBS(m))$ is the likelihood that we can assess using Laplace method and $p(\theta|WBS(m))$ the prior distribution. Then, the expected value and the variance of θ_l are obtained by:

$$\begin{cases} \mathbb{E}(\theta_l|\mathbf{d}, \mathbf{z}) = \sum_{m=1}^{4} \hat{\theta}_l^m p(WBS(m)|D) \\ \mathbb{V}(\theta_l|\mathbf{d}, \mathbf{z}) = \sum_{m=1}^{4} [\mathbb{V}(\theta_l|\mathbf{d}, \mathbf{z}, WBS(m)) + (\hat{\theta}_l^m)^2] p(WBS(m)|\mathbf{d}, \mathbf{z}) - \mathbb{E}(\theta_l|\mathbf{d}, \mathbf{z})^2 \end{cases}$$

where $\hat{\theta}_l^m = \mathbb{E}(\theta_l|\mathbf{d}, \mathbf{z}, WBS(m))$.

2.4 | Model comparison

In our simulations, we compare the performance of the WBS method to the classical Bayesian SSVS and the frequentist Lasso. Our evaluation criteria are bias, pointwise prediction performance, the posterior distribution of the prediction performance, and variable selection performance.

Bias

The bias of the estimator $\widehat{\theta}_l$, $l \in \{1, ..., L\}$, is defined as the difference $Bias(\widehat{\theta}_l) = \mathbb{E}(\widehat{\theta}_l) - \theta_l$ between the expected value of the estimator and the true values of the coefficient θ_l, where $\widehat{\theta}_l$ is the posterior mean of θ_l for the Bayesian methods and the Lasso estimates otherwise.

Performance of prediction

To evaluate the prediction performance of each method, we use the root mean square error (RMSE), computed as the sample standard deviation of the differences between predicted and observed values:

$$RMSE(\widetilde{\widehat{\mathbf{d}}}) = \sqrt{MSE(\widetilde{\widehat{\mathbf{d}}})} = \sqrt{\mathbb{E}((\widetilde{\widehat{\mathbf{d}}} - \widetilde{\mathbf{d}})^2)} \approx \sqrt{\frac{1}{n}\frac{1}{K}\sum_{i=1}^{n}\sum_{i=1}^{K}(\widetilde{\widehat{d}_{i,k}} - \widetilde{d}_{i,k})^2}$$

where $\widetilde{\mathbf{d}}$ are observed doses for patients in the validation dataset and $\widetilde{\widehat{\mathbf{d}}} = \widehat{\theta}_0 + \widehat{\theta}^T \widetilde{\mathbf{Z}}$ are the estimations given by the model for these doses. Additional details are presented in A.

Bayesian predictive model accuracy

Unlike frequentist methods, the Bayesian approach provides a posterior distribution, which provides a basis for computing the log pointwise predictive density, proposed by Gelman[25]. We adapt this approach to the repeated measures structure of the Irinotecan data set. Denoting the prior distribution as $p(\boldsymbol{\theta}, \mathbf{I}, \boldsymbol{\gamma}, \sigma_\gamma^2, \sigma_\epsilon^2)$ and the posterior distribution as $p_{post}(\boldsymbol{\theta}, \mathbf{I}, \boldsymbol{\gamma}, \sigma_\gamma^2, \sigma_\epsilon^2 \mid data)$, the posterior predictive distribution of $\widetilde{\mathbf{d}}$ is

$$p_{post}(\widetilde{\mathbf{d}} \mid data) = \int p(\widetilde{\mathbf{d}}|\boldsymbol{\theta}, \mathbf{I}, \boldsymbol{\gamma}, \sigma_\gamma^2, \sigma_\epsilon^2) p_{post}(\boldsymbol{\theta}, \mathbf{I}, \boldsymbol{\gamma}, \sigma_\gamma^2, \sigma_\epsilon^2 \mid data) d\boldsymbol{\theta}\, d\mathbf{I}\, d\boldsymbol{\gamma}\, d\sigma_\gamma^2\, d\sigma_\epsilon^2.$$

In practice, the distribution $f_{i,k}(\mathbf{d})$ depends on $(\boldsymbol{\theta}, \mathbf{I}, \boldsymbol{\gamma}, \sigma_\gamma^2, \sigma_\epsilon^2)$, which are not known. We thus work with the log pointwise predictive density:

$$\text{LPD} = \sum_{i=1}^{n}\sum_{k=1}^{K} \log p_{post}(\widetilde{d}_{i,k}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \log \int p(\widetilde{d}_{i,k}|\boldsymbol{\theta}, \mathbf{I}, \boldsymbol{\gamma}, \sigma_\gamma^2, \sigma_\epsilon^2) p_{post}(\boldsymbol{\theta} \mid data) d\boldsymbol{\theta}\, d\mathbf{I}\, d\boldsymbol{\gamma}\, d\sigma_\gamma^2\, d\sigma_\epsilon^2.$$

To assess the predictive accuracy for the $n \times K$ data points in a dataset, the expected log pointwise predictive density (ELPD) is defined as follows:

$$\text{ELPD} = \sum_{i=1}^{n}\sum_{k=1}^{K} E_{f_{i,k}}\{\log p_{post}(\widetilde{d}_{i,k})\}$$

where $f_{i,k}(\mathbf{d})$ is the distribution representing the true data-generating process for $d_{i,k}$.

For computational details, see B. Because each model has the same number of parameters, and they are estimated in the same way, one might simply compare their best-fit log predictive densities directly. We thus select the model with the smallest value of −LPD.

Precision of variable selection

A method's variable selection ability may be assessed through its false positive rate (FPR) and false negative rate (FNR). The FPR is defined to be the number of non-predictive variables, i.e., having true coefficient 0, that are wrongly selected by the method (false positive = FP) divided by the total number of actual non-predictive variables (N),

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

where TN is the number of true negatives, i.e., covariates having true coefficient 0. In the Bayesian framework, a variable is considered selected when its posterior median is equal to one. For Lasso methods, all non-selected covariates are estimated as exactly zero. Note that $FPR = 1 - TNR$, where TNR is the true negative rate, or specificity.

Similarly, the false negative rate (FNR) is defined as

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP}$$

where FN is the number of false negatives, TP is the number of true positives and $P = FN + TP$ is the total number of positive selections. Additionally, $FNR = 1 - TPR$ where TPR is the true positive rate, or sensitivity.

3 | SIMULATION DESIGN

We simulated a sample of patients treated for colorectal metastatic cancer by Irinotecan, to imitate the structure of our motivating data set. The protocol is based on a theoretical dose of 180 mg/m^2. Severity levels of toxicities were simulated with grades taking on integer values from 0 (indicating no toxicity of that type) to 4 (the most severe level), or to 3 for nausea and asthenia.

3.1 | Covariate distributions used in the simulations

Numerical parameter choices for the simulations were based on case study covariate distributions (cf. Table 4). $X_{..1}$ and $X_{..2}$ are assumed to be fixed in time. All other variables change over time and were simulated with an autocorrelation structure of order 1 (ρ) between cycles according to a Gaussian copula based procedure.

3.2 | Dose generation

The dose for individual $i \in \{1, ..., n\}$ at cycle $k \in \{1, ..., K\}$ was generated from the model

$$d_{i,k} = D_{max} - \boldsymbol{\beta}^T \boldsymbol{z}_{i,k} + \epsilon_{i,k}$$

where $\epsilon_{i,k} \sim \mathcal{N}(0, \sigma^2)$. Four realistic scenarios were simulated: two involving clinician 1's clinical relevance weights (scenarios 1a and 1b) and two involving clinician 2's clinical relevance weights (scenarios 2a and 2b). First, we selected variables with clinical relevance weights greater than specific thresholds; then, we set coefficients $\boldsymbol{\beta}$ of the selected variables as a linear function

TABLE 3 Simulation parameters for the covariates.

Variables	Original variables	Probability distribution	Parameters
$x_{..1}$	Age $\geq$ 80 years	$\mathcal{N}(\mu, s^2)$	$\mu = 63$, $s = 12$
$x_{..2}$	Treatment line 3, > 3	$\mathcal{G}eom(r_1)$	$r_1 = 0.5$
$x_{..3}$	Weight loss > 10%	$\mathcal{B}(r_2)$	$r_2 = 0.1$
$x_{..4}$	WHO score	$\mathcal{B}(l, r_3)$	$l = 4$, $r_3 = 0.5$
$x_{..5}$	Bilirubin >35 μmol/L	$\mathcal{B}(r_3)$	$r_4 = 0.1$
$x_{..6}$	Vomiting	$\mathcal{G}eom(r_5)$ truncated to s_1	$r_5 = 0.2$, $s_1 = 4$
$x_{..7}$	Nausea	$\mathcal{G}eom(r_6)$ truncated to s_2	$r_6 = 0.2$, $s_2 = 3$
$x_{..8}$	Diarrhea	$\mathcal{G}eom(r_7)$ truncated to s_3	$r_7 = 0.2$, $s_3 = 4$
$x_{..9}$	Asthenia	$\mathcal{G}eom(r_8)$ truncated to s_4	$r_8 = 0.2$, $s_4 = 3$
$x_{...,10}$	Neutropenia	$\mathcal{G}eom(r_9)$ truncated to s_5	$r_9 = 0.2$, $s_5 = 4$
$x_{...,11}$	Thrombopenia	$\mathcal{G}eom(r_{10})$ truncated to s_6	$r_{10} = 0.2$, $s_6 = 4$
$x_{...,12}$	Anemia	$\mathcal{G}eom(r_{11})$ truncated to s_7	$r_{11} = 0.2$, $s_7 = 4$

of the clinical relevance weights such that all doses were included between D_{min} and D_{max}. We constructed also an unrealistic scenario where the clinical relevance weights used to simulate the scenario are taken equal to $100 - \mathbf{w}'$ for clinician 1. Thus, for example, vomiting of grade 4 is linked to a smaller clinical weight that vomiting of grade 3. So, considering priors built on clinician expertise could favor inclusion of wrong variables. Values of the thresholds and coefficients $-\boldsymbol{\beta}$ are given in Table 5 .

To assess the extent to which the prior deviates from the true parameter value, we calculate Spearman's rank correlation coefficients between:

- Corrected clinical relevance weights used to build prior distributions ($\mathbf{w'^c}$), that is clinical relevance weights with minimal and maximal thresholds to 20 and 80 respectively;
- Weights used to build scenarios ($\mathbf{w'^u}$), that is only clinical relevance weights greater than thresholds specific to each scenario are non-zeros weights.

It allows one to measure the extent to which, as one variable increases, the other variable tends to increase. Table 6 gives Spearman coefficient between each model and each scenario.

For each of the five scenarios, we studied three cases: ($n = 10, K = 3$), ($n = 10, K = 5$) and ($n = 20, K = 5$). Values of parameters were derived from case study covariates (cf. Table 7). We set $\sigma = 5$ based on suggestions reported in the literature.

For each of the fifteen sub-scenarios, 1000 data sets were simulated. Each data set consisted of a training set of size $n \times K$, along with an independent validation set of size 50×5. For each simulated data set, we fit eight models on the training data sets using the Lasso method, the classical SSVS method, and the WBS method with clinical relevance weights provided by each of the four clinicians, with mixture and with BMA. All analyses were performed using *R* software 3.3.2 version and packages *glmmLasso* and *R2jags*. In *R2jags*, *autojags* was implemented with 3 chains and a burn-in of 1000 and, as a convergence criterion, Gelman and Rubin's potential scale reduction factor $Rhat = 1.1$.

3.3 | Prior parameter settings

In Bayesian estimation of mixed models (cf. subsection 2.2, equation 1), prior distributions for the parameters $\boldsymbol{\theta}$, σ_γ and σ_ϵ must be defined. After a sensitivity analysis in which several parameter values were tested and model performance was evaluated in a few scenarios, an inverse-gamma distribution with shape and scale parameters (1, 1) and (0.2, 1), respectively, were chosen for the covariance of the random effects σ^2_γ and the residual variance σ^2_ϵ, (cf. Table 8). For $\boldsymbol{\theta}$, see subsection 2.2.

TABLE 4 Values of $-\boldsymbol{\beta}$ used for each simulation scenario.

Variable	Realistic scenarios				Unrealistic scenario
	1a	1b	2a	2b	A
$X_{..1}$	18.6	13.4	14.4	12.6	0
$X_{..2}$	0,0	0,6.70	0,0	0,0	0,0
$X_{..3}$	0	6.70	0	4.19	0
$X_{..4}$	0,0,0,0	0,0,0,0	0,0,9.63,24.1	0,0,8.39,21.0	18.8,15.1,15.1,15.1
$X_{..5}$	18.6	13.4	9.63	8.39	0
$X_{..6}$	0,0,0,0	0,0,10.7,12.1	0,0,16.9,24.1	0,6.29,14.7,21.0	18.8,15.1,0,0
$X_{..7}$	0,0,0	0,0,10.7	0,0,12.0	0,0,10.5	18.8,15.1,0
$X_{..8}$	0,0,0,18.6	0,0,10.7,13.4	0,0,12.0,24.1	0,4.19,10.5,21.0	18.8,0,0,0
$X_{..9}$	0,0,18.6	0,6.70,13.4	0,0,9.63	0,0,8.39	17.0,0,0
$X_{...,10}$	0,0,18.6,18.6	0,9.39,13.4,13.4	0,0,0,12.0	0,0,6.29,10.5	18.8,0,0,0
$X_{...,11}$	0,18.6,18.6,18.6	0,13.4,13.4,13.4	0,0,0,0	0,0,4.19,6.29	0,0,0,0
$X_{...,12}$	0,0,0,18.6	0,6.70,10.7,13.4	0,0,0,0	0,0,4.19,6.29	18.8,0,0,0
Number of variables actually used	10	20	11	19	12
Weights' threshold	100	50	40	20	80

TABLE 5 Spearman's rank correlation coefficient between corrected clinical relevance weights (in lines) and weights used to build scenarios (in columns)

WBS	Scenarios				
	1 a	1 b	2 a	2 b	A
1	0.668	0.948	0.426	0.585	-0.857
2	0.379	0.505	0.913	0.949	-0.318
3	0.520	0.424	0.562	0.669	-0.300
4	0.436	0.610	0.655	0.788	-0.454

TABLE 6 Parameter values in the simulations.

Parameter	Values
n	10 or 20
K	3 or 5
J	12
L	35
ρ	0.5
D_{min}	50
D_{max}	180
σ	5
S	0.02
W_{max}	100

The models used by the SSVS and WBS algorithms also require specification of fixed prior hyper-parameters, $\boldsymbol{\tau}^2$ and $\boldsymbol{g\tau}^2$, which we based on the data (cf. equation 2). For $l \in \{1, ..., L\}$, τ_l should be chosen such that if $\theta_l \sim \mathcal{N}(0, \tau_l^2)$, then θ_l can be "safely" replaced by 0. Moreover, g_l (>1) should be chosen such that if $\theta_l \sim \mathcal{N}(0, g_l\tau_l^2)$, then a non-0 estimate of θ_l should be included in the final model. Some suggestions can be found in[11]. Because the smallest numerical coefficient value for our simulated data is 4.19 (cf. Table 5), we take $3\tau_l$ to be equal to the maximum value at which θ_l would be equivalent to 0. In the following, based on our sensitivity analysis, we assume $\tau_l = 1$ and $g_l = 100^2$.

TABLE 7 Choice of prior distributions.

Parameter	Prior distribution	Hyper-parameters
σ^2_γ	$Inverse - Gamma(\alpha_1, \beta_1)$	$\alpha_1 = 1, \beta_1 = 1$
σ^2_ϵ	$Inverse - Gamma(\alpha_2, \beta_2)$	$\alpha_2 = 0.2, \beta_2 = 1$
θ_0	$N(\mu_0, s_0^2)$ truncated to 0	$\mu_0 = 100, s_0 = 100$

4 | SIMULATION RESULTS

4.1 | Bias

The coefficients' bias (cf. Figures **??** and **??** in supplementary material) are on average negative and dramatically decrease with their real values; therefore, the methods underestimate the coefficients of variables actually used. For realistic scenarios, the coefficients of variables actually not used are estimated to be no more than -3 mg/m^2, that is -1.67 % of the baseline dose.

4.2 | Prediction performance

Tables 9 and 10 summarize the average prediction performance (RMSEs) for validation sets over 1 000 runs, along with the precision of selected variables (mean FPR and mean FNR) for each method. Figures 3 show the distribution of RMSE for all scenarios with $n = 10$ and $K = 3$.

Recall that the weights of clinicians 1 and 2 were used for scenarios 1a and 1b and scenarios 2a and 2b, respectively, in the simulations. For scenarios 1a and 1b, WBS1 has a slightly smaller RMSE than that of SSVS and the WBS2, WBS3 and WBS4 when $n = 10$ patients and $K = 3$ cycles (cf. Table 9 and Figures 3 a and 3 b). However, the improvement is so small that only a mean improvement of less than 2.3 mg/m^2 in the RMSE (for a baseline dose of 180 mg/m^2) compared with the SSVS method is observed. The performance of WBS4 is similar to that of WBS1 as the Spearman's rank correlation coefficient between scenarios 1a/1b and WBS(m) takes the highest values for WBS1 and WBS4 (cf. Table 6). WBS with the mixture does not succeed to improve prediction performance, contrary to WBS with BMA that manages inconsistent weighting by the different physicians by favoring WBS1 and WBS4 models.

For scenarios 2a and 2b, the performance is approximately the same for all methods considered (cf. Table 9 and Figures 3 c, 3 d). The performance of WBS is similar to that of SSVS, regardless of the weights used for the prior. For $n = 10$ patients with $K = 3$ cycles, the best performance is not achieved with WBS2, as might have been expected, but rather is best with the WBS4 and WBS with BMA. The WBS4's RMSE decreases by approximatively 1.7 mg/m^2 (for a baseline dose of 180 mg/m^2), when $n = 10$ patients and $K = 3$ cycles, compared with that of the SSVS method.

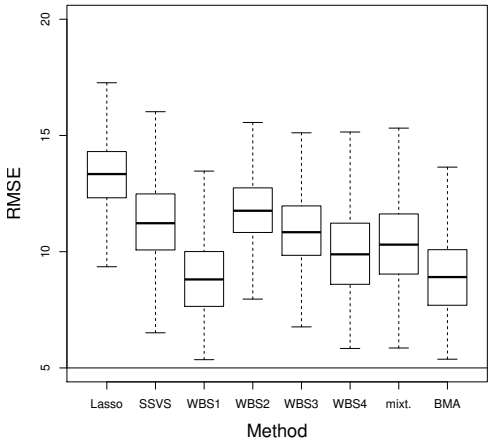
The scenario A is the unrealistic scenario built using the importance weights {100 - clinical relevance weights of clinician 1}. For this scenario, WBS models perform poorly and the smallest RMSE is obtained for SSVS model (cf. Table 10 and Figure 3 e). Although all Spearman coefficients are negative between scenario A and WBS(m), $m \in \{1, 2, 3, 4\}$, the higher Spearman coefficients for this scenario are obtained for WBS3 and WBS2 models for which we find the best performances among the WBS models (cf. Table 6).

Table **??** and Figure **??** in supplementary material present the $-1\times$ mean log pointwise predictive density over 1000 replications for each method. The log pointwise predictive density between the different methods is similar, as is the case for RMSE.

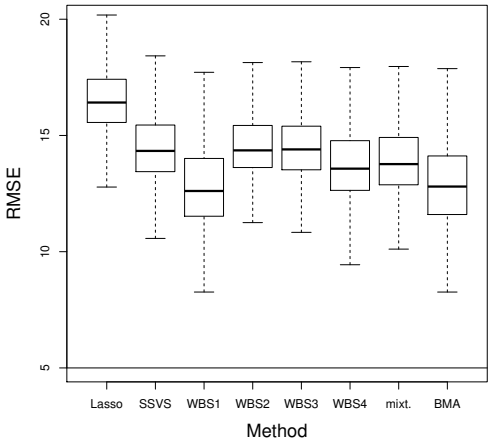
In general, the frequentist LASSO shows worse performance except for the unrealistic scenario. Performance improves with an increase in sample size and increased number of cycles. The WBS method predicts no better than the other methods for $n = 20$. The RMSE is close to the variance of the error term ($\sigma = 5$) for all methods.

4.3 | Precision of selected variables

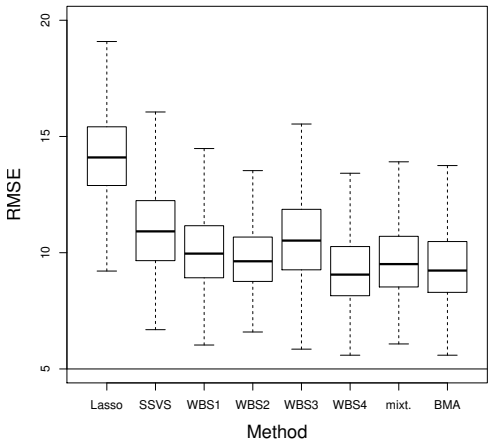
Figure 4 and Figures **??** and **??** in supplementary material respectively show the percentage of times in which each variable is selected by the model, the true positive rate plotted against the false positive rate and the percentage of times versus real



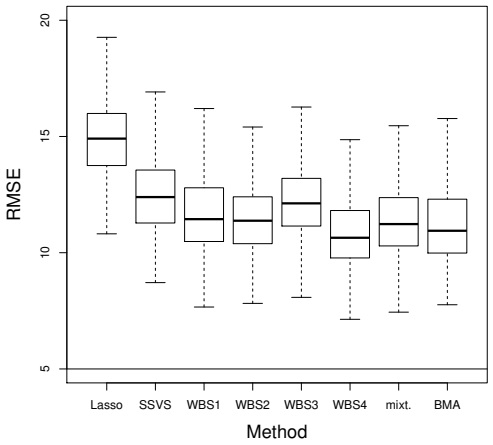
(a) Scenario 1a: $(n, K) = (10, 3)$



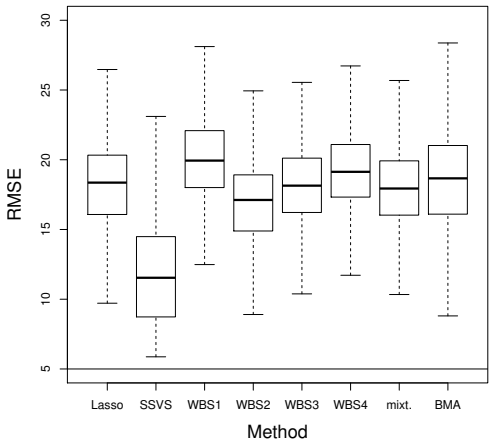
(b) Scenario 1b: $(n, K) = (10, 3)$



(c) Scenario 2a: $(n, K) = (10, 3)$



(d) Scenario 2b: $(n, K) = (10, 3)$



(e) Scenario A: $(n, K) = (10, 3)$

FIGURE 3 Boxplot of root mean square error for each method and for each scenario with $n = 10$ and $K = 3$

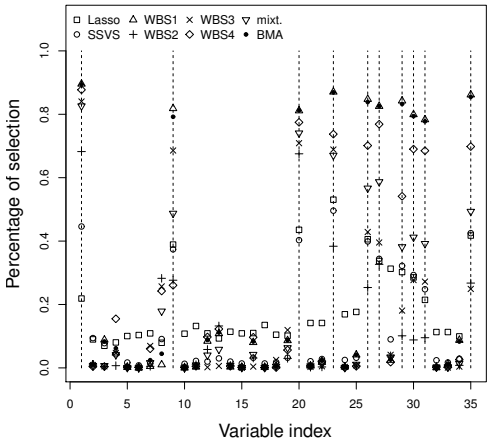
coefficients' values when $n = 10$ and $K = 3$. For realistic scenarios, the WBS method performs better than the other methods in terms of variable selection, with both lower FPRs and FNRs in general. The SSVS and WBS methods select a few variables not actually used (less than 5 % FPR), but they fail more often to correctly select variables actually used, with up to 84 % FNR. For small patient samples, the Lasso shows very large FNRs. For scenarios 1a,1b, 2a and 2b, the FNRs of the Lasso are 65 %, 81 %, 67 % and 79 %, respectively. When the number of observations increases, the Lasso wrongly selects many variables, with FPRs of 45 %, 85 %, 56 % and 80 %, while the SSVS and WBS methods only select the true variables with FPR and FNR values close to 0.

Again, for scenarios 1a and 1b, the WBS1 model gives the best performance in selecting variables, followed by WBS with BMA and WBS4 (cf. Table 9 , Figures 4 a, 4 b and Figures **??**, **??**, **??**, **??** in supplementary material). Indeed, WBS1 improves the selection ability, with 46 % and 30 % smaller FNR in scenarios 1a and 1b, respectively, compared with the SSVS method, when $n = 10$ and $K = 3$. In the simulations, recall that clinical relevance weights were used (1) to choose what variables are selected and (2) to estimate the coefficients associated with the variables used for dose reduction. Therefore, the ideal clinical relevance weights used in the models should all be either 0 or 100, implying that clinicians are either sure to use a covariate or not in his/her decision process. Scenario 1a is an example in which all covariates that have an influence on doses are associated with a clinical relevance weight of 100 for the WBS1 method. For scenarios 2a and 2b, like WBS with BMA, the WBS4 and WBS1 models give low FPR and FNR because of their higher clinical relevance weights (cf. Table 9 , Figures 4 c, 4 d and Figures **??**, **??**, **??**, **??** in supplementary material). In these two scenarios, the WBS4 model yields 29 % and 58 % FNR, respectively, when $n = 10$ and $K = 3$, while the SSVS method yields much larger FNR values 58 % and 79 %. Overall, WBS1 and WBS4 are associated with larger clinical relevance weights than WBS2 and WBS3 are (cf. sum of clinical relevance weights for each clinician in Table 1), and they select a higher number of covariates. Interestingly, the WBS model that selects the highest number of covariates actually used, and a smaller number of covariates actually not used, is not necessarily the one that uses the clinical relevance weights of the clinician on which the simulation setting is based.

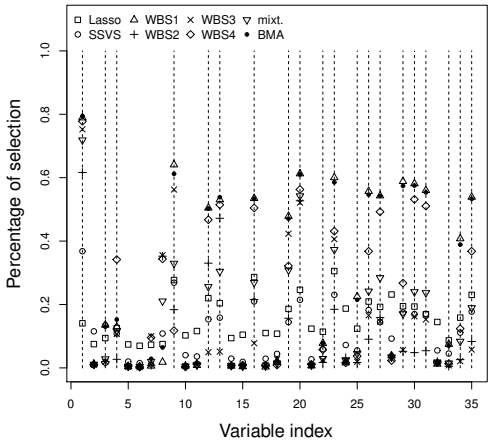
Furthermore, WBS1 fails to select a higher percentage of variables for scenario 1b, with an FNR of 54 % for $(n, K) = (10, 3)$, compared with scenario 1a, with an FNR of 16 % for $(n, K) = (10, 3)$), generated based on fewer variables and larger coefficients for the variables' effect on selected doses. Similarly, scenario 2a has a lower FNR than scenario 2b. Finally, for $n = 20$ patients, the FNR values are close to 0 for scenario 2a but are high for scenario 2b. As expected, the ability to correctly select variables increases with increasing sample size and number of cycles.

For scenario A, the variable selection performance of our method is as bad as the predictive performance for this scenario (cf. Table 10 , Figures 4 e and Figures **??**, **??**, in supplementary material).

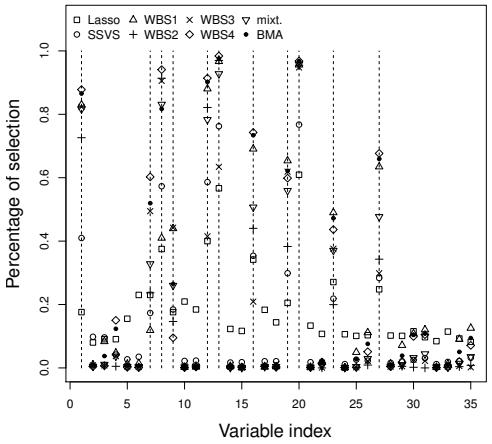
The Lasso failed to converge in 1 to 11 % of the simulated cases, depending on the scenario, with a median of 6 %, while the other methods failed to converge in less than 2 % of the cases.



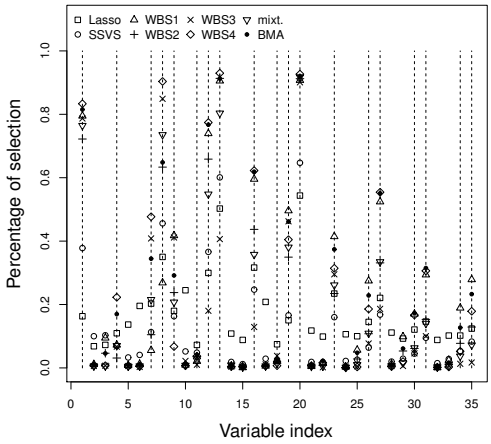
(a) Scenario 1a: $(n, K) = (10, 3)$



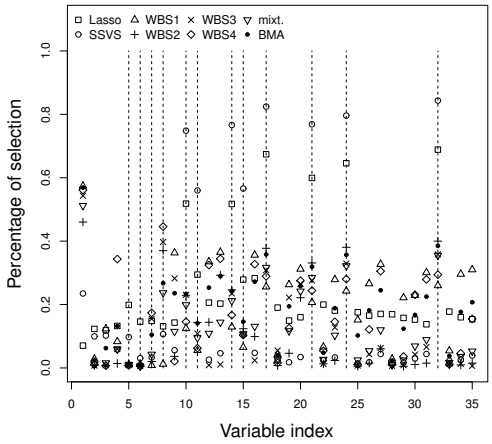
(b) Scenario 1b: $(n, K) = (10, 3)$



(c) Scenario 2a: $(n, K) = (10, 3)$



(d) Scenario 2b: $(n, K) = (10, 3)$



(e) Scenario A: $(n, K) = (10, 3)$

FIGURE 4 Percentage of times where each variable is selected by the model for scenarios with $n = 10$ and $K = 3$. Variables actually used are identified by dashed lines.

TABLE 8 Simulation results for realistic scenarios - Mean root mean square errors and mean false positive rate and mean false negative rate over 1 000 replications.

	Lasso	Classical SSVS	WBS					
			1	2	3	4	mixture	BMA
Scenario 1a: $(n, K) = (10, 3)$								
RMSE	13.4 (1.58)	11.3 (1.83)	8.99 (1.77)	11.8 (1.57)	10.9 (1.65)	10.0 (1.94)	10.4 (1.85)	9.11 (1.87)
FPR	12	3	3	3	3	4	2	*3*
FNR	65	62	16	68	53	33	45	*17*
Scenario 1a: $(n, K) = (10, 5)$								
RMSE	9.46 (2.00)	7.32 (1.46)	6.53 (0.907)	8.07 (1.87)	7.15 (1.38)	6.78 (1.05)	6.84 (1.17)	6.59 (0.920)
FPR	37	1	1	1	1	1	1	*1*
FNR	23	13	1	22	9	4	5	*1*
Scenario 1a: $(n, K) = (20, 5)$								
RMSE	6.17 (0.674)	5.50 (0.558)	5.63 (0.563)	5.63 (0.574)	5.63 (0.577)	5.64 (0.578)	5.62 (0.572)	5.63 (0.565)
FPR	45	0	0	0	0	0	0	*0*
FNR	2	1	0	0	0	0	0	*0*
Scenario 1b: $(n, K) = (10, 3)$								
RMSE	16.5 (1.46)	14.6 (1.82)	12.9 (1.94)	14.6 (1.55)	14.6 (1.62)	13.8 (1.93)	14.0 (1.74)	13.1 (2.11)
FPR	11	5	1	4	4	4	3	*2*
FNR	81	84	54	84	81	66	76	*55*
Scenario 1b: $(n, K) = (10, 5)$								
RMSE	13.1 (2.14)	12.9 (1.41)	9.79 (1.54)	13.3 (1.23)	13.2 (1.25)	11.4 (1.54)	12.2 (1.44)	9.81 (1.56)
FPR	42	3	0	2	3	2	2	*0*
FNR	41	73	33	76	73	50	63	*33*
Scenario 1b: $(n, K) = (20, 5)$								
RMSE	6.92 (0.774)	6.33 (0.708)	6.20 (0.624)	6.78 (1.10)	6.75 (1.03)	6.34 (0.688)	6.37 (0.693)	6.22 (0.626)
FPR	85	0	0	0	0	0	0	*0*
FNR	3	9	7	15	14	9	10	*7*
Scenario 2a: $(n, K) = (10, 3)$								
RMSE	14.1 (1.84)	11.1 (2.00)	10.2 (1.83)	9.79 (1.54)	10.6 (1.92)	9.34 (1.72)	9.72 (1.74)	9.54 (1.83)
FPR	12	3	4	0	1	2	1	*3*
FNR	67	58	36	44	44	29	38	*29*
Scenario 2a: $(n, K) = (10, 5)$								
RMSE	9.50 (1.74)	7.62 (1.27)	7.22 (1.11)	7.38 (1.12)	7.19 (1.20)	7.03 (1.00)	7.12 (1.08)	6.97 (1.01)
FPR	39	1	1	0	0	1	0	*1*
FNR	24	23	11	19	12	11	13	*8*
Scenario 2a: $(n, K) = (20, 5)$								
RMSE	6.37 (0.704)	5.72 (0.566)	5.70 (0.559)	5.71 (0.556)	5.67 (0.559)	5.72 (0.559)	5.68 (0.553)	5.68 (0.544)
FPR	56	0	0	0	0	0	0	*0*
FNR	3	2	1	2	1	2	1	*0*
Scenario 2b: $(n, K) = (10, 3)$								
RMSE	14.9 (1.66)	12.6 (1.82)	11.8 (1.84)	11.5 (1.64)	12.3 (1.72)	10.9 (1.77)	11.5 (1.72)	11.3 (1.92)
FPR	12	4	2	1	1	1	1	*2*
FNR	79	79	61	68	72	58	68	*58*
Scenario 2b: $(n, K) = (10, 5)$								
RMSE	11.2 (1.97)	9.84 (1.42)	9.00 (1.25)	9.54 (1.19)	9.61 (1.43)	8.70 (1.15)	9.13 (1.25)	8.70 (1.17)
FPR	39	2	1	0	0	0	0	*1*
FNR	45	63	47	60	57	46	54	*44*
Scenario 2b: $(n, K) = (20, 5)$								
RMSE	6.96 (0.730)	6.65 (0.664)	6.40 (0.621)	6.83 (0.659)	6.62 (0.638)	6.49 (0.624)	6.57 (0.634)	6.39 (0.605)
FPR	80	0	0	0	0	0	0	*0*
FNR	8	31	23	37	32	26	30	*23*

Standards deviations are shown in parentheses.

TABLE 9 Simulation results for unrealistic scenario - Mean root mean square errors and mean false positive rate and mean false negative rate over 1 000 replications.

	Lasso	**Classical SSVS**	**WBS**					
			1	**2**	**3**	**4**	**mixture**	**BMA**
Scenario A: $(n, K) = (10, 3)$								
RMSE	18.2 (3.26)	11.9 (3.78)	20.2 (3.53)	16.8 (3.46)	18.0 (3.61)	19.3 (3.69)	17.9 (3.53)	18.6 (4.53)
FPR	16	5	23	6	8	16	9	*18*
FNR	60	49	88	78	79	82	82	*79*
Scenario A: $(n, K) = (10, 5)$								
RMSE	10.3 (2.27)	6.97 (0.872)	8.47 (2.60)	7.45 (1.69)	7.38 (1.84)	7.75 (2.18)	7.59 (1.78)	7.40 (1.72)
FPR	46	2	6	3	3	4	3	*4*
FNR	23	24	36	29	27	28	30	*26*
Scenario A: $(n, K) = (20, 5)$								
RMSE	6.92 (0.678)	5.72 (0.567)	6.06 (0.663)	5.77 (0.614)	5.68 (0.553)	5.71 (0.555)	5.78 (0.598)	5.69 (0.546)
FPR	70	1	2	1	1	1	1	*1*
FNR	10	5	16	9	6	7	9	*6*

Standards deviations are shown in parentheses.

TABLE 10 Case study - Description of covariates.

Variables	0	1	2	3	4
Age $\geq$ 80 years	19 (58)	14 (42)	-	-	-
Weight loss > 10%	32 (97)	1 (3)	-	-	-
WHO score	7 (21)	21 (64)	5 (15)	-	-
Bilirubin >35 μmol/L	33 (100)	-	-	-	-
Treatment line 3, > 3	33 (100)	-	-	-	-
Toxicity grades					
Vomiting	33 (100)	-	-	-	-
Nausea	15 (45)	16 (48)	2 (6)	-	-
Diarrhea	17 (52)	15 (45)	1 (3)	-	-
Asthenia	4 (12)	18 (55)	10 (30)	1 (3)	-
Neutropenia	28 (85)	3 (9)	2 (6)	-	-
Thrombopenia	28 (85)	5 (15)	-	-	-
Anemia	23 (70)	9 (27)	1 (3)	-	-

Percentages are shown in parentheses.

5 | CASE STUDY

We extracted data for patients treated for metastatic colorectal cancer with a combination of drugs including Irinotecan from the Georges Pompidou University Hospital (HEGP) I2B2 warehouse with IRB approval. The protocol encompasses a theoretical dose of 180 mg/m^2 and a theoretical cycle of chemotherapy of 14 days. All patients who were treated with any other protocol including Irinotecan were excluded. One cycle was defined as doses being given over one to three successive days. When a patient received a dose more than 28 days later, we considered that to be the beginning of another protocol. The data included the covariates age, weight, total bilirubin, WHO score, treatment line, and the seven toxicity types registered on each cycle of chemotherapy, as described in section 3, using the physicians' elicited clinical relevance weights.

In our database, we found 185 patients who had data for the first $K = 5$ cycles, among whom 70 patients had at least one cycle with complete toxicity data available. To test our method with a small sample, we randomly selected $n = 10$ patients having $N = 33$ complete forms. Among the 10 included patients, only 3 toxicities of grade 3 (asthenia) were observed (see Table 11 for description of covariates). For the case study, RMSEs were computed over these patients.

Results are shown in Table 12 . The best performance was obtained by WBS3, with the smallest RMSE and 2 variables selected : age > 80 years and asthenia of grade 3, respectively linked to dose reductions of -13.4 mg/m^2 and -85.3 mg/m^2. Asthenia 3 is selected by all methods, except the Lasso, and age is selected by these same methods, except WBS2. The clinical relevance weight linked to age is 100 for WBS1, WBS3, and WBS4 but is only 60 for WBS2. Only WBS2 yields worse performance than the classical SSVS, with only one variable selected. The Lasso selects six variables. WBS1, WBS4, WBS with mixture and WBS with BMA choose the same variables. WBS3 and WBS4 select fewer variables than WBS1 and WBS4, possibly because of smaller sums of weights (cf. Table 1). Other selected variables are anemia grade 2, selected by three methods (coefficient < -10); asthenia grade 2 and thrombopenia grade 1, only selected by the Lasso and SSVS; and WHO score and anemia grade 1, only selected by the Lasso. The WBS methods do not select thrombopenia grade 1, likely because this toxicity has clinical relevance weights of only 40 and 0.

In general, the WBS method produces coherent results, with WBS likely to correctly select ordinal variables having the highest clinical value, while the Lasso appears to select variables almost completely randomly, for instance, selecting asthenia grade 2 while not selecting asthenia grade 3.

TABLE 11 Case study results - Estimated coefficients, root mean square errors and - log pointwise predictive density.

Variables	Weights of clinicians: 1	2	3	4	Lasso	Classical SSVS	WBS with prior from clinicians: 1	2	3	4	Mixture	BMA
Interecept	-	-	-	-	178	174	176	173	178	174	175	176
Age ≥ 80 years	100	60	100	80	-11.1*	-3.21	-10.1*	-5.56	-13.4*	-8.79*	-10.2*	-11.0*
Weight loss > 10 %	50	20	50	80	0	-0.267	-0.664	-0.177	-0.737	-1.72	-0.625	-0.850
WHO score 1	0	0	0	0	-10.9*	0.268	-0.00462	0.0618	0.194	0.0936	0.0663	0.0895
WHO score 2	20	0	0	20	0	-0.721	-0.454	-0.590	-0.754	-0.438	-0.522	-0.570
Nausea 1	0	0	0	10	0	-0.337	-0.342	-0.605	-2.16	-1.95	-0.925	-1.31
Nausea 2	20	10	10	30	0	0.483	0.305	2.32	2.24	1.89	1.35	1.38
Diarrhea 1	0	0	0	0	0	-0.784	-1.44	-0.399	-2.34	-0.268	-0.714	-1.55
Diarrhea 2	40	20	50	20	0	1.23	0.313	0.222	0.175	0.180	0.191	0.235
Asthenia 1	10	10	0	10	0	-0.238	0.0724	0.0915	0.0812	0.0534	0.0955	0.0733
Asthenia 2	50	10	0	50	7.39*	-6.57*	-2.54	-1.61	-3.36	-1.58	-1.88	-2.65
Asthenia 3	100	40	70	70	0	-86.7*	-82.1*	-80.5*	-85.3*	-81.9*	-81.3*	-83.2*
Neutropenia 1	0	0	0	0	0	0.357	0.331	0.303	0.190	0.640	0.224	0.328
Neutropenia 2	70	0	0	20	0	-0.956	-1.26	-3.39	-0.764	-1.64	-1.56	-1.21
Thrombopenia 1	40	0	0	0	7.11*	-7.65*	-2.78	-2.39	-2.39	-2.81	-2.42	-2.62
Anemia 1	0	0	0	0	-3.09*	0.445	-0.0263	-0.0508	0.0188	0.0380	-0.0572	0.000632
Anemia 2	50	0	0	20	-38.3*	-5.99	-14.2*	-7.35	-1.93	-13.6*	-10.7*	-9.21*
Number of selected variables					6	3	3	1	2	3	3	3
RMSE					25.2	17.6	17.3	18.5	17.1	17.5	17.5	17.3
- Log pointwise predictive density					-	142	141	144	141	142	143	

* represents the variables selected by the model.

6 | DISCUSSION

In this paper, we have proposed the WBS (Weights-based SSVS) method derived from SSVS (Stochastic Search Variable Selection). The WBS method uses elicited clinical relevance weights to construct prior distributions for covariate coefficient inclusion probabilities in a regression model for longitudinal clinical practice data. An extensive simulation study showed that, as long as clinicians do not provide absurd expertise, compared with the classical SSVS method, the WBS method exhibited better performance for all criteria considered (RMSE, log pointwise predictive density) and produced lower rates of both false positives and false negatives. As expected, performance improved with increasing sample size. WBS performance depended not only on the covariates' importance weights but also on the weights' sum, which must be calibrated carefully to obtain good performance. In our simulation study, we chose a large number of covariates implicated in dose reduction, and therefore, models with lower prior weights showed poorer performance because of low variable selection rates. The Lasso showed poor performance compared with WBS and SSVS, confirming that Bayesian methods outperform frequentist methods when working with small samples.

To our knowledge, variable selection methods incorporating informative prior distributions have been considered mainly in genomic settings. In this context, priors are established by exploiting an expansive literature on genetic variant severity, both from experimental and bioinformatic points of view. Therefore, in genetic settings, variability of priors is considered to be low, and simulation studies of performance variability due to prior specification have received little interest [19]. This is not the case for covariates retrieved from EHR, for which the literature is extremely limited regarding their respective effects on disease severity and thus their importance in medical decision making.

In our setting, only binary variables were used. We dichotomized continuous variables because it is more straightforward for clinicians to provide elicited weights for binary variables. The proposed method allows for the use of continuous variables, however. Another limitation of our formulation is that no hierarchical constraints were imposed in variable selection; that is, if one variable was included, nothing forced another to be included. For instance, if vomiting grade 3 is included in the model, vomiting of grade 4 should also be included. This limitation could be overcome by using the Farcomeni approach developed for SSVS models [14]. Additionally, in the Irinotecan data and similar settings, one may force all estimated regression coefficients to be negative or null because associated variables should imply either a dose reduction or no dose adjustment. Furthermore, this work began the analysis only from the second cycle, as we focused on toxicities. However, the first cycle can be included by considering all toxicities in this cycle to be null.

Many additional elaborations are possible. In this study, although the visit times for different patients may differ substantially, depending on the side effects generated in the previous cycle, we did not incorporate time or covariates of previous cycles. In our model, toxicities are not dependent variables but rather are covariates, and if there is a trend with respect to time, it should have already be summarized by the toxicities. In clinical practice, oncologists consider all of the patient's treatment history in choosing the dose and the next visit time. Furthermore, toxicities depend directly on previous doses and times. One possible extension might be to adapt this method to bivariate models in a dynamic treatment regime framework. Moreover, our model does not take into account the probability of changing the dose but only "how to reduce the dose when the clinician should to do it". Another extension of our model could be a conditional model on "when to change".

Concerning elicitation of clinical relevance weights, our case study results suggest that clinicians may not have an accurate perception of their actual decision process for dose adjustment. Indeed, most individuals are influenced by cognitive, psychological, and emotional factors that often prevent decision makers from choosing the best rational option available [26]. For example, in clinical practice, if a physician observes major side effects in a patient one day, his/her emotions and actual experience can lead him/her to reduce the dose of the next patient, even if common sense says to maintain the current dose.

The elicited scores differed between clinicians, as may be expected. When analyzing real data, one solution could be to combine the physicians' experiences to build more informative priors either by applying a mathematical aggregation rule, such as using a mixture prior with the physicians weighted equally, or by allowing the experts to interact with each other to obtain a consensus prior [27,28]. This interaction may be face-to-face or may involve exchanges of information without direct contact. The prevailing mathematical approaches are averaging and pooling [29,30]. Other such approaches deviate from these traditional approaches by treating the elicited information as data. In this paper, our proposal is to use Bayesian Model Averaging that permits to combine the results obtained by WBS method for each clinicians' set of weights by giving more importance to the most performant model.

Finally, modeling medical decision making may be regarded as a first step to modeling the complex relationships between covariates, dose reduction decisions, and survival. Our methodology may be extended to focus on dose combinations in frail patients, with the goal to optimize survival. Finally, our method may also be applied in other settings in which we wish to account

for experts' opinion when selecting specific variables in small samples. Indeed, after choosing what variables have to be include in the model and categorizing them, clinical relevance weights may be elicited from many clinicians for each variable. Then, to combine the different sets of clinical weights, a mixture prior can be used after fitting the models for each clinicians' weights or directly BMA as suggested in the paper. With this approach, the methodology is generally applicable to other settings having this data structure

ACKNOWLEDGMENTS

The authors thank Dr. Aziz Zaanan, Dr. Céline Lepère, Dr. Simon Pernot and Dr. Anne-Laure Pointet from Georges Pompidou European Hospital for their help with elicitation. The authors thank Angelika Geroldinger as well for her useful suggestions. The simulations were performed at the HPCaVe at UPMC-Sorbonne Université.

AUTHOR CONTRIBUTIONS

Anne-Sophie Jannot and Sarah Zohar made equal contributions and are co-last authors.

FUNDING

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: A part of this project was supported by French National Cancer Institut (INCa) grant INCA_10801. Moreno Ursino was funded by INCa grant INCA_9539.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

How to cite this article: Sandrine Boulet, Moreno Ursino, Peter Thall, Anne-Sophie Jannot, and Sarah Zohar (???), Bayesian variable selection based on clinical relevance weights in small sample studies - Application to colon cancer, *Stat Med.*, *???*.

APPENDIX

A CONVERGENCE OF RMSE

Take the simple case of the linear regression model:

$$d_i = \theta_0 + \boldsymbol{\theta}^T \mathbf{z}_i + \epsilon_i,\ \epsilon_i\ iid \sim \mathcal{N}(0, \sigma^2),\ i \in \{1, ..., n\}$$

Once coefficients $(\theta_0, \boldsymbol{\theta}^T)$ have been estimated, we predict the dose d_i by

$$\widehat{d}_i = \widehat{\theta}_0 + \widehat{\boldsymbol{\theta}}^T \mathbf{z}_i$$

Because $MSE(\widehat{\mathbf{d}}) = RMSE(\widehat{\mathbf{d}})^2 = \mathbb{E}((\mathbf{d} - \widehat{\mathbf{d}})^2)$, the MSE can be estimated by

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{i=1}^{n}(d_i - \widehat{d}_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}((\theta_0 - \widehat{\theta}_0) + (\boldsymbol{\theta}^T - \widehat{\boldsymbol{\theta}}^T)\mathbf{z}_i + \epsilon_i)^2 \end{aligned}$$

In the ideal case in which $(\widehat{\theta}_0, \widehat{\boldsymbol{\theta}}^T) = (\theta_0, \boldsymbol{\theta}^T)$, we obtain

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n}\left(\frac{\epsilon_i}{\sigma}\right)^2$$

Now, $\left(\frac{\epsilon_i}{\sigma}\right)_{1\leq i\leq n}$ are independent, standard normal random variables; thus, the sum of their squares is distributed according to a chi-squared distribution with n degrees of freedom:

$$\sum_{i=1}^{n}\left(\frac{\epsilon_i}{\sigma}\right)^2 \sim \chi_n^2$$

We deduce the expectation and the variance:

$$\mathbb{E}(MSE) = \sigma^2,\ \mathbb{V}(MSE) = 2 \times \frac{\sigma^4}{n}$$

B PRACTICAL COMPUTATION OF LOG POINTWISE PREDICTIVE DENSITY

To compute the LPD in practice, it is possible to evaluate the expectation using draws from $p_{post}(\boldsymbol{\theta}, \mathbf{I}, \boldsymbol{\gamma}, \sigma_\gamma^2, \sigma_\epsilon^2)$, the usual posterior simulations, which are labeled $\boldsymbol{\theta}^s, \mathbf{I}^s, \boldsymbol{\gamma}^s, c^{2s}, \sigma_\epsilon^{2s}$, $s = 1, ..., S$:

$$\widehat{\text{LPD}} = \sum_{i=1}^{n}\sum_{k=1}^{K}\log\left(\frac{1}{S}\sum_{s=1}^{S}p(d_{i,k}|\boldsymbol{\theta}^s, I^s, \boldsymbol{\gamma}^s, c^{2s}, \sigma_\epsilon^{2s})\right)$$

The log pointwise predictive density $\widehat{\text{LPD}}$ is the sum over patients and cycles of the log of the mean over MCMC iterations of the probability that a new dose $\widetilde{d}$ would be obtained by the estimated model.

The following steps are suggested to compute the LPD in our case:

1. For $s = 1, ..., S$, sample $(\boldsymbol{\theta}^s, c^{2s}, \sigma_\epsilon^{2s})$ from $p_{post}(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\gamma^2, \sigma_\epsilon^2)$:
   $(\boldsymbol{\theta}^s, c^{2s})$ are the values obtained for the s^{th} iteration of the MCMC;
2. For $s = 1, ..., S$, for $i = 1, ..., n$, draw γ_i^s from $\mathcal{N}(0, c^{2s})$;
3. Compute the probability that $p(\widetilde{\mathbf{d}}|\boldsymbol{\theta}^s) = (\sigma_\epsilon^{2s})^{-\frac{nk}{2}}\exp\left(-\frac{(\widetilde{\mathbf{d}} - \theta_0^s - \mathbf{z}\boldsymbol{\theta}^s - \boldsymbol{\gamma}^s)^T(\widetilde{\mathbf{d}} - \theta_0^s - \mathbf{z}\boldsymbol{\theta}^s - \boldsymbol{\gamma}^s)}{2\sigma_\epsilon^{2s}}\right)$.

References

1. Frankovich Jennifer, Longhurst Christopher A., Sutherland Scott M.. Evidence-Based Medicine in the EMR Era. *New England Journal of Medicine.* 2011;365(19):1758–1759.
2. Tenenbaum Jessica D., Avillach Paul, Benham-Hutchins Marge, et al. An informatics research agenda to support precision medicine: seven key areas. *Journal of the American Medical Informatics Association.* 2016;23(4):791–795.
3. Bailey Peter S. J., Chang David K., Nones Katia, et al. *Genomic analyses identify molecular subtypes of pancreatic cancer.* 2016.
4. Tibshirani Robert. Regression shrinkage and selection via the lasso: a retrospective: Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2011;73(3):273–282.

5. Fan Jianqing, Li Runze. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association.* 2001;96(456):1348–1360.

6. Efron Bradley, Hastie Trevor, Johnstone Iain, Tibshirani Robert. Least Angle Regression. *The Annals of Statistics.* 2004;32(2):407–451.

7. Zou Hui, Hastie Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2005;67(2):301–320.

8. Bondell Howard D., Reich Brian J.. Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics.* 2008;64(1):115–123.

9. O'Hara R. B., SillanpÃďÃď M. J.. A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis.* 2009;4(1):85–117.

10. Mitchell T. J., Beauchamp J. J.. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association.* 1988;83(404):1023–1032.

11. George Edward I., McCulloch Robert E.. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association.* 1993;88(423):881–889.

12. Kuo Lynn, Mallick Bani. Variable Selection for Regression Models. *SankhyÄĄ: The Indian Journal of Statistics, Series B (1960-2002).* 1998;60(1):65–81.

13. Ishwaran Hemant, Rao J. Sunil. Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics.* 2005;33(2):730–773.

14. Farcomeni Alessio. Bayesian Constrained Variable Selection. ;:28.

15. Zhang Lin, Baladandayuthapani Veerabhadran, Mallick Bani K., et al. Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society. Series C, Applied statistics.* 2014;63(4):595–620.

16. Hobert James P., Casella George. The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association.* 1996;91(436):1461–1473.

17. Park Trevor, Casella George. The Bayesian Lasso. *Journal of the American Statistical Association.* 2008;103(482):681–686.

18. Chipman Hugh, George Edward I., McCulloch Robert E.. The Practical Implementation of Bayesian Model Selection. In: Beachwood, OH: Institute of Mathematical Statistics 2001 (pp. 65–116).

19. Kitchen Christina M. R., Weiss Robert E., Liu Gang, Wrin Terri. HIV-1 viral fitness estimation using exchangeable on subsets priors and prior model selection. *Statistics in Medicine.* 2007;26(5):975–990.

20. Thall Peter F., Ursino Moreno, Baudouin VÃľronique, Alberti Corinne, Zohar Sarah. Bayesian Treatment Comparison Using Parametric Mixture Priors Computed from Elicited Histograms. *Statistical methods in medical research.* 2017;:962280217726803.

21. Manceau Gilles, Imbeaud Sandrine, ThiÃľbaut RaphaÃńle, et al. Hsa-miR-31-3p Expression Is Linked to Progression-free Survival in Patients with KRAS Wild-type Metastatic Colorectal Cancer Treated with Anti-EGFR Therapy. *Clinical Cancer Research.* 2014;20(12):3338–3347.

22. Jansman Frank G.A., Sleijfer Dirk T., Coenen Jules L.L.M., De Graaf Jacques C., Brouwers Jacobus R.B.J.. Risk Factors Determining Chemotherapeutic Toxicity in Patients with Advanced Colorectal Cancer:. *Drug Safety.* 2000;23(4):255–278.

23. Bekele B. Nebiyou, Thall Peter F.. Dose-Finding Based on Multiple Toxicities in a Soft Tissue Sarcoma Trial. *Journal of the American Statistical Association.* 2004;99(465):26–35.

24. Raftery Adrian E., Madigan David, Hoeting Jennifer A.. Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association.* 1997;92(437):179–191.

25. Gelman Andrew, Hwang Jessica, Vehtari Aki. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*. 2014;24(6):997–1016.

26. Gorini Alessandra, Pravettoni Gabriella. An overview on cognitive aspects implicated in medical decisions. *European Journal of Internal Medicine*. 2011;22(6):547–553.

27. Winkler Robert L.. The Consensus of Subjective Probability Distributions. *Management Science*. 1968;15(2):B–61.

28. Clemen Robert T., Winkler Robert L.. Combining Probability Distributions From Experts in Risk Analysis. *Risk Analysis*. 1999;19(2):187–203.

29. Genest Christian, McConway Kevin J.. Allocating the weights in the linear opinion pool. *Journal of Forecasting*. 1990;9(1):53–73.

30. Genest Christian, Zidek James V.. Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science*. 1986;1(1):114–135.