A new modeling approach for quantifying expert opinion in the drug discovery process

Ariel Alonso[1] **Elasma Milanzi**[2] **Geert Molenberghs**[1,2]
Christophe Buyck[3] **Luc Bijnens**[3]

[1] *I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

[2] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[3] *Janssen Pharmaceutica, Johnson & Johnson, B-2340 Beerse, Belgium*

Abstract

Expert opinion plays an important role when choosing clusters of chemical compounds for further investigation. Often, the process by which the clusters are assigned to the experts for evaluation, the so-called selection process, and the qualitative ratings given by the experts to the clusters (chosen/not chosen) need to be jointly modeled to avoid bias. This approach is referred to as the joint modeling approach. However, misspecifying the selection model may impact the estimation and inferences on parameters in the rating model, which are of most scientific interest. We propose to incorporate the selection process into the analysis by adding a new set of random effects to the rating model and, in this way, avoid the need to model it parametrically. This approach is referred to as the *combined* model approach. Through simulations, the performance of the combined and joint models were compared in terms of bias and confidence interval coverage. The estimates from the combined model were nearly unbiased and the derived confidence intervals had coverage probability around 95% in all scenarios considered. In contrast, the estimates from the joint model were severely biased under some form of misspecification of the selection model and fitting the model was often numerically challenging. The results show that the combined model may offer a safer alternative on which to base inferences when there are doubts about the validity of the selection model. Importantly, thanks to its greater numerical stability, the combined model may outperform the joint model even when the latter is correctly specified.

Keywords: Combined model, Selection bias, Shared parameter, Sensitivity.

1 Introduction

Developing chemical compounds into effective drugs is an expensive and lengthy process. Therefore, pharmaceutical companies need to carefully evaluate the amount of evidence supporting their potential, before investing more resources on them [1]. Expert opinion is a valuable tool for the assessment of these compounds at early stages in the drug discovery process [2, 3]. In practice, similar compounds are grouped into clusters that are qualitatively assessed by experts regarding their selection for future scrutiny. Further, with appropriate statistical methods, these assessments can be quantified as a success probability for each cluster, where success is defined as being chosen for further investigation [4, 5].

The large number of clusters typically involved in these studies implies that a selection procedure, by which every expert chooses or gets assigned a number of clusters for evaluation, needs to be implemented. Alonso *et al.* [6] showed that some selection procedures may introduce a selection bias in the rating process and lead to invalid conclusions. In these scenarios, complex joint hierarchical models, describing the selection and rating processes, are required to get valid results. Actually, these authors demonstrated that, even in the absence of selection bias in the rating process, one often needs to jointly model both processes to get valid estimates. Ideally, one would like to know all factors influencing the selection process beforehand. However, in practice, such information is seldom available and making assumptions on the selection process is then virtually inescapable.

We shall consider two approaches to account for the selection process. In the first approach, two generalized linear mixed models (GLMM) are used to describe the rating and selection processes and it is assumed that, given some random effects, both processes are independent. We shall refer to this approach as the joint modeling approach. The joint modeling approach is closely related to the shared parameter (SP) and generalized shared parameter (GSP) modeling frameworks, used to describe a Missing Not At Random (MNAR) mechanism in missing data analysis [7, 8]. In addition, the assumption of conditional independence is at the core of complex hierarchical models developed to describe, for instance, the joint evolution of longitudinal and survival outcomes and, in the present work, it simplifies the joint distribution of the rating and selection processes, facilitating the joint fit of both models [9–11].

This approach hinges on the assumption that the distribution for the selection process is correctly specified. In general, if the selection model is misspecified then the estimates of the parameters in the rating model may be biased and inferential procedures, like obtaining confidence intervals, may be affected as well. Therefore, a sensitivity analysis to assess the stability of the results is always highly recommended [12].

Our second approach is based on the so-called combined model introduced by Booth *et al.* [13] and Molenberghs *et al.* [14] for members of the exponential family, where an extra set of random effects is used to account for overdispersion in correlated outcomes. Similarly,

in this work, we propose to take into account the selection process by adding a new set of random effects to the rating model. We extensively study the performance of both approaches using theoretical elements and simulations. Our results show that the combined model could be a robust alternative to the joint model when analyzing this type of data, even when the selection model is correctly specified. Therefore, we think that the combined model may serve two purposes: (i) it may be a reliable tool for sensitivity analysis and (ii) when there are doubts regarding the performance of the joint model, it may be a safer alternative on which to base inferences.

The paper is organized as follows; before presenting the two modeling approaches in Sections 3 and 4, respectively, we discuss the motivating case study in Section 2. The simulation study is presented in Section 5 followed by the analysis of the case study in Section 6. Brief concluding remarks in Section 7 wind up the paper.

2 Case Study

The pharmaceutical company Johnson&Johnson carried out a study to evaluate the potential of $22\,015$ clusters of chemical compounds, in order to determine those that warrant further screening. In total, 147 experts were asked to evaluate several of these clusters and their assessments were coded as 1 if they recommended the cluster for further screening, -1 if not recommended and 0 if indifferent. The response was dichotomized for the analysis. We adopted a coding scheme where 1 corresponds to a positive recommendation and 0 otherwise. However, the methodology presented can easily accommodate other coding schemes as well.

Experts carried out the evaluation of the clusters using the desk-top application Third Dimension Explorer (3DX) [15]. In a regular session a random subset of clusters, selected from the entire set of $22\,015$, was assigned to each expert for evaluation. Clusters were presented with additional information that included their size, the structure of some of their distinctive members like the compound with the lowest/highest molecular weight, and 1–3 other randomly chosen members of the clusters. The application was designed to support multiple sessions that would allow the experts to stop and resume the evaluation at their own convenience. A new random subset of clusters, excluding the ones already rated, was assigned for evaluation only when all the clusters in the previous subset were evaluated, or when the experts resumed the evaluation after interrupting the previous session for a break. Clusters assigned but not evaluated could, in principle, be assigned again in another session.

The histogram in Figure 1 displays the distribution of the number of clusters evaluated by the experts. Clearly, the distribution is positively skewed, indicating that, as one would expect, many experts opted to evaluate few clusters. Indeed, 25% of the experts evaluated less than 345 clusters, 50% less than 1200, and 75% of the experts evaluated less than 2370 clusters. Evidently, the large differences in the number of clusters evaluated by the experts are not

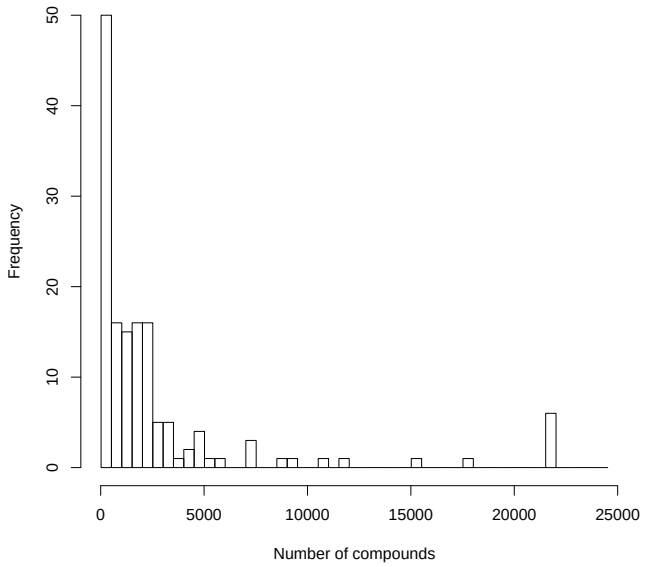

Figure 1: *Histogram for the number of clusters rated by the experts. The height of a bar indicates the number of experts whose number of rated clusters fall within the range given by the width of the bar.*

the result of the random allocation, but rather are dictated by the number of evaluation sessions each expert found convenient. Actually, the possibility of interrupting and resuming the evaluation session at will allowed the experts to influence the selection process and, hence, standard models that assume complete randomization may not be appropriate.

Alonso *et al.* [6] explored how such a design may lead to biased results and discussed a method for correcting the problem. Basically, these authors carried out two different analyses: One that completely ignored the selection process and another one that accounted for it using the joint modeling approach. The results from these two analyses were staggeringly different. These differences and the information available about the study design clearly call for a cautious analysis of these data.

3 The Joint Modeling Approach

To facilitate the decision making process, Milanzi [4], Milanzi *et al.* [5] and Alonso *et al.* [6] proposed to summarize the large number of qualitative assessments given by the experts into a single probability of success for every cluster. Denoting by $\boldsymbol{Y}_i = (Y_{ij})_{j\in\Lambda_i}$ the vector of ratings associated with expert i, where Λ_i is the subset of all clusters evaluated by the expert

and $i = 1, \ldots, n$, these authors considered the following logistic-normal model

$$\text{logit}\left[P\left(Y_{ij} = 1|b_i\right)\right] = \beta_j + b_i, \tag{1}$$

where $Y_{ij} = 1$ if expert i recommends cluster j for further scrutiny, β_j is a fixed parameter characterizing the effect of cluster C_j with $j \in \Lambda_i$ and the b_is are independent expert effects with $b_i \sim N(0, \sigma_b^2)$. Based on model (1), the marginal probability of success for cluster C_j can be calculated by integrating over the random effect, i.e.,

$$P\left(Y_j = 1\right) = \int \frac{\exp\left(\beta_j + b_i\right)}{1 + \exp\left(\beta_j + b_i\right)} \phi(b_i|0, \sigma_b^2)\, db_i, \tag{2}$$

where $\phi(b_i|0, \sigma_b^2)$ denotes a normal density with mean zero and variance σ_b^2.

Notice that the likelihood associated with model (1) suffers from a severe dimensionality problem. Indeed, the vector of fixed effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)^T$ has dimension $N = 22\,015$ and the dimension (N_i) of the response vector $\boldsymbol{Y}_i$ ranges from 20 to $22\,015$. As a consequence, serious computational issues can emerge when fitting model (1) with the most commonly available computing resources. Milanzi *et al.* [5] developed a procedure that allows to handle these issues with a very small loss of efficiency and in the present work the dimensionality problem will just be discussed briefly for the case study analysis.

Alonso *et al.* [6] pointed out that model (1) quantifies the probability that expert i would rate cluster j as 1, given that he actually evaluates it and introduced two GLMM $P\left(X_{ij} = x_{ij}|a_i, \alpha_j\right)$ and $P\left(Y_{ij} = y_{ij}|X_{ij} = x_{ij}, b_i, \beta_j\right)$ to describe the selection and rating procedures, respectively, where $X_{ij} = 1$ if expert i evaluates cluster j and 0 otherwise. The vectors of expert-specific random effects $(a_i, b_i)^T$ are assumed to follow a bivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$.

These authors stated that there is selection bias in the rating process if the rating that would be given to a cluster by an expert depends on whether he selects it or not for evaluation, i.e., if $P\big(Y_{ij} = y_{ij}|X_{ij} = 1, b_i\big) \neq P\big(Y_{ij} = y_{ij}|X_{ij} = 0, b_i\big)$ and showed that absence of selection bias in the rating process is equivalent to the validity of the following conditional independence assumption

$$P\left(Y_{ij} = y_{ij}, X_{ij} = x_{ij}|a_i, b_i\right) = P\left(Y_{ij} = y_{ij}|b_i\right) P\left(X_{ij} = x_{ij}|a_i\right). \tag{3}$$

Essentially, (3) states that for every expert the rating and selection procedures are independent and governed by different, although possibly marginally correlated, random effects. In the most general scenario, the potential of cluster j can be quantified as

$$P\left(Y_j = 1\right) = \int\int P\left(Y_{ij} = 1|a_i, b_i\right)\, \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma})\, da_i db_i, \tag{4}$$

where $\phi(\cdot|\mathbf{0}, \boldsymbol{\Sigma})$ denotes a bivariate normal density with mean zero and covariance matrix $\boldsymbol{\Sigma}$ and

$$P(Y_{ij} = 1|a_i, b_i) = E_X\left[P(Y_{ij} = 1|X_{ij} = x_{ij}, b_i)\right] \tag{5}$$
$$= P(Y_{ij} = 1|X_{ij} = 1, b_i)\, P(X_{ij} = 1|a_i) + P(Y_{ij} = 1|X_{ij} = 0, b_i)\, P(X_{ij} = 0|a_i)\,.$$

Clearly, there is information about how the experts rated the clusters they evaluated and, therefore, $P(Y_{ij} = 1|X_{ij} = 1, b_i)$ can be estimated from the data using, for instance, model (1). Furthermore, there is also information about which clusters every expert evaluated and this information could be used to estimate $P(X_{ij} = 1|a_i)$. However, the events $\{Y_{ij} = y_{ij}|X_{ij} = 0, b_i\}$ are counterfactual and we do not have information about how the experts would have rated a cluster they did not evaluate if, contrary to fact, they had evaluated it. As a result, the probabilities $P(Y_{ij} = 1|X_{ij} = 0, b_i)$ are not identifiable from the data without additional assumptions.

Importantly, under conditional independence, one has that $P\big(Y_{ij} = y_{ij}|X_{ij} = 1, b_i\big) = P\big(Y_{ij} = y_{ij}|X_{ij} = 0, b_i\big)$ and (4) simplifies to (2). Alonso *et al.* [6] argued that, even in absence of a selection bias in the rating process, the selection process given by $P(X_{ij} = x_{ij}|a_i, \beta_j, \alpha_j)$ may need to be explicitly modeled. The reason for this counter intuitive finding is that, even though the selection procedure does not affect the ratings, ignoring it when constructing the likelihood function may induce bias in the estimates of β_j, σ_b^2 and, consequently, in the estimate of $P(Y_j = 1)$. To illustrate this important fact, in the rest of this section, we will assume conditional independence and that the components of the vectors $\boldsymbol{Y}_i, \boldsymbol{X}_i \in \{0, 1\}^N$ are also independent conditionally on the random effects, with $\boldsymbol{X}_i$ denoting the vector of selection-indicators for expert i.

The parameters of interest are estimated based on the complete data $\{\boldsymbol{Y}_i, \boldsymbol{X}_i\}$. The vector of ratings can be decomposed as $\boldsymbol{Y}_i = (\boldsymbol{Y}_{0i}^T, \boldsymbol{Y}_{1i}^T)^T$, where $\boldsymbol{Y}_{1i} \in \{0, 1\}^{N_i}$ is the sub-vector associated with the clusters the expert actually evaluated, $\boldsymbol{Y}_{0i}^T$ is the obvious complement and $N_i = \mathbf{1}^T\boldsymbol{X}_i$. Alonso *et al.* [6] showed that, under conditional independence, the marginal likelihood takes the form

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \prod_i^n P(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}, \boldsymbol{X}_i = \boldsymbol{x}_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})\,, \tag{6}$$

where

$$P(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}, \boldsymbol{X}_i = \boldsymbol{x}_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$$
$$= \int\int P(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta})\, P(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\alpha})\; \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma})\, da_i db_i \tag{7}$$

and

$$\begin{aligned} P(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta}) &= \prod_{j \in \Lambda_i} P(Y_{1ij} = y_{1ij}|b_i, \beta_j), \\ P(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\alpha}) &= \prod_{j}^{N} P(X_{ij} = x_{ij}|a_i, \alpha_j). \end{aligned}$$

Using the maximum likelihood estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\sigma}_b^2$ obtained from (6), one can estimate the probabilities of success by substituting $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}_b^2$ into (2). Note, however, that to estimate $\boldsymbol{\beta}$, σ_b^2, one may need to explicitly model the selection process using, for example, GLMM. An important special instance where the selection mechanism can be ignored is when the selection and rating processes are also marginally independent, i.e, when $\phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}) = \phi(a_i|0, \sigma_a^2)\phi(b_i|0, \sigma_b^2)$ and have a disjoint parametric space. In fact, under these assumptions (7) simplifies to

$$\begin{aligned} & P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}, \boldsymbol{X}_i = \boldsymbol{x}_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2\right) \\ &= \int P\left(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\alpha}\right)\, \phi(a_i|0, \sigma_a^2)\, da_i \int P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta}\right)\, \phi(b_i|0, \sigma_b^2)\, db_i. \end{aligned}$$

Consequently, regarding the parameters of interest $\boldsymbol{\beta}$ and σ_b^2, the contribution of expert i to the likelihood becomes

$$\int P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta}\right)\, \phi(b_i|0, \sigma_b^2)\, db_i, = \int \left[\prod_{j \in \Lambda_i} P\left(Y_{1ij} = y_{1ij}|b_i, \beta_j\right)\right]\, \phi(b_i|0, \sigma_b^2)\, db_i.$$

The previous expression is the contribution of expert i to the likelihood when the selection mechanism has been discarded. Therefore, in this scenario, if conditional independence holds, the selection procedure can be ignored. This setting will result, for instance, if a fully random allocation of the clusters to experts is implemented, so that the experts have no influence whatsoever on the selection process. However, if experts can influence the selection process then a selection model may need to be incorporated into the analysis in order to achieve valid results, even when there is not a selection bias in the rating process.

4 Combined Model Approach

In this section, a new modeling framework for quantifying expert opinion will be introduced. To this end, let us assume that there exists independent latent selection traits θ_{ij} for every expert-cluster combination. Further, we will denote by $f\left(y_{ij}, \theta_{ij}, b_i\right)$ the distribution of the vector $(y_{ij}, \theta_{ij}, b_i)^T$ and it will be assumed that, conditional on $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iN})^T$ and b_i, the components of $\boldsymbol{Y}_i$ are independent. More specifically, it will be assumed that

$P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i | b_i, \boldsymbol{\theta}_i\right) = \prod_j^N P\left(Y_{ij} = y_{ij} | b_i, \theta_{ij}\right)$. Basically, the latter assumption states that conditional on the selection traits, the ratings of expert i are independent. Similarly, it will be assumed that the random variables θ_{ij} and b_i are independent as well. Under all the previous assumptions one has

$$\begin{aligned} f\left(\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{\theta}_i, b_i\right) =& P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i | b_i, \boldsymbol{\theta}_i\right) f\left(\boldsymbol{\theta}_i\right) f(b_i) \\ =& \left[\prod_j^N P\left(Y_{ij} = y_{ij} | b_i, \theta_{ij}\right) f\left(\theta_{ij}\right)\right] f(b_i). \end{aligned} \tag{8}$$

In expression (8), $P\left(Y_{ij} = y_{ij} | b_i, \theta_{ij}\right)$ describes the rating process conditional on the latent selection trait and the expert effect b_i. It is important to point out that, although θ_{ij} and b_i are marginally independent, conditional on Y_{ij} they are dependent.

The rating and selection processes are not independent if $P\left(Y_{ij} = y_{ij} | b_i, \theta_{ij}\right) \neq P\left(Y_{ij} = y_{ij} | b_i\right)$. Essentially, unlike in the joint model where the association between the selection and rating processes is implicitly captured by the correlation between a_i and b_i, in the combined model this association is explicitly given in $P\left(Y_{ij} = y_{ij} | b_i, \theta_{ij}\right)$.

The new model is completed by making parametric assumptions for the distributions in (8). For practical reasons that will become clear later we have chosen

$$Y_{ij} | b_i, \theta_{ij} \sim \text{Bernoulli}(\theta_{ij}\pi_{ij}), \qquad \pi_{ij} = \frac{\exp(\beta_j + b_i)}{1 + \exp(\beta_j + b_i)},$$

$$\theta_{ij} \sim f(\theta_{ij}) = \text{Beta}\left(\theta_{ij} | \lambda, \tau\right), \quad b_i \sim N(0, \sigma_b^2).$$

In this framework, the probability of success for compound C_j is given by

$$P\left(Y_j = 1\right) = \int\int P\left(Y_{ij} = 1, \theta_{ij}, b_i\right) \, d\theta_{ij} \, db_i = \frac{\lambda}{\lambda + \tau} E_b\left(\pi_{ij}\right). \tag{9}$$

Some insight can be gained by assuming that a larger selection trait is associated with a higher probability of selection. Under this assumption, clusters evaluated have a higher probability of being chosen for further investigation than unevaluated ones. Indeed, to fix ideas let us assume that $X_{ij} = 1$ if $\theta_{ij} \geq \gamma_{ij}$ and zero otherwise, where the γ_{ij}s are the threshold values at which the latent selection traits are manifested. It can be easily shown that (details in the

web supporting materials)

$$P\left(Y_{ij}=1|X_{ij}=1,b_i\right) = \pi_{ij}\frac{\int_{\gamma_{ij}}^{1}\theta_{ij}f(\theta_{ij})\,d\theta_{ij}}{\int_{\gamma_{ij}}^{1}f(\theta_{ij})\,d\theta_{ij}}, \tag{10}$$

$$P\left(Y_{ij}=1|X_{ij}=0,b_i\right) = \pi_{ij}\frac{\int_{0}^{\gamma_{ij}}\theta_{ij}f(\theta_{ij})\,d\theta_{ij}}{\int_{0}^{\gamma_{ij}}f(\theta_{ij})\,d\theta_{ij}}.$$

Using some properties of the beta and the incomplete beta distributions one can show that, as expected, $P\left(Y_{ij}=1|b_i,X_{ij}=0\right)\leq P\left(Y_{ij}=1|b_i,X_{ij}=1\right)$ if $\gamma_{ij}\in(0,1)$. Alonso *et al.* [6] called this inequality the monotonicity assumption and showed that, when there is a selection bias in the rating process and monotonicity holds, the use of likelihood (6) in combination with (2) will produce an upper bound for the probabilities of success. The flexibility of the combined model allows to accommodate monotone and non-monotone settings, however, the validity of the results obtained from it relies on several untestable assumptions, like the multiplicative effect of θ_{ij} on π_{ij} and the use of a convenient conjugate distribution for θ_{ij}. Some additional insights into the properties and interpretation of the combined model are provided in the web supporting materials.

Considering the previously introduced partition $\boldsymbol{Y}_i=(\boldsymbol{Y}_{0i}^T,\boldsymbol{Y}_{1i}^T)^T$ and the corresponding counterpart $\boldsymbol{\theta}_i=(\boldsymbol{\theta}_{0i}^T,\boldsymbol{\theta}_{1i}^T)^T$, expression (8) takes the form

$$f\left(\boldsymbol{Y}_{0i},\boldsymbol{Y}_{1i},\boldsymbol{\theta}_{0i},\boldsymbol{\theta}_{1i},b_i\right)=P\left(\boldsymbol{Y}_{0i}|\boldsymbol{\theta}_{0i},b_i\right)P\left(\boldsymbol{Y}_{1i}|\boldsymbol{\theta}_{1i},b_i\right)f\left(\boldsymbol{\theta}_{0i},\boldsymbol{\theta}_{1i}\right)f(b_i),$$

and after marginalizing over the subvectors $\boldsymbol{Y}_{0i}$, $\boldsymbol{\theta}_{0i}$ one gets

$$f\left(\boldsymbol{Y}_{1i},\boldsymbol{\theta}_{1i},b_i\right)=P\left(\boldsymbol{Y}_{1i}|\boldsymbol{\theta}_{1i},b_i\right)f\left(\boldsymbol{\theta}_{1i}\right)f(b_i).$$

The parameter estimates are derived using the marginal likelihood obtained after integrating out the random effects b_i and $\boldsymbol{\theta}_{1i}$. This process is carried out in two steps. First, after analytically integrating over $\boldsymbol{\theta}_{1i}$ the likelihood contribution for each expert follows as

$$\begin{aligned} L_c^*(\boldsymbol{\beta},\lambda,\tau,b_i) &= \int f\left(\boldsymbol{Y}_{1i},\boldsymbol{\theta}_{1i},b_i\right)\,d\boldsymbol{\theta}_{1i}, \\ &= \prod_{j=1}^{N_i}\left\{\frac{1}{1+k}(\pi_{ij})^{y_{ij}}\left[(1-\pi_{ij})+k\right]^{1-y_{ij}}\right\}, \end{aligned} \tag{11}$$

where $k=\tau/\lambda$ and, eventually, in the second step the marginal likelihood can be obtained by numerically integrating over the normal random effect b_i, using readily available statistical

software, i.e., the parameter estimates follow from maximizing

$$L_m(\boldsymbol{\beta}, k, \sigma^2) = \prod_i^n \int L_c^*(\boldsymbol{\beta}, \lambda, \tau, b_i)\phi(b_i|0, \sigma^2)\, db_i. \tag{12}$$

Estimation of k instead of individual parameters λ and τ is necessary to avoid identification problems [14]. For example, if individual parameters were to be estimated, the following sets of parameters would lead to the same solution for (12): $(\lambda = 2, \tau = 1)$, $(\lambda = 4, \tau = 2)$, and $(\lambda = 6, \tau = 3)$. The multiplicative factor in (9) becomes, $1/(1 + k)$.

In both this model as well as in the joint model, the connection between rating and selection processes is at the level of latent variables. In the joint model, this is via the correlation between the random effects a_i and b_i in (3). In the combined model, the link follows from the latent variable θ_{ij} and its threshold γ_{ij}, as in (10). Thus, the connection with the corresponding missing-data concepts is at the level of the likelihood.

5 Simulation Study

To numerically evaluate the performance of the combined and joint models a simulation study was designed. The data were generated by mimicking the case study introduced in Section 2. This notwithstanding, the size of the simulated data sets were chosen so that all models could be fitted using maximum likelihood in order to minimize the numerical noise and provide a clearer idea regarding the performance of both approaches. Two hundred data sets were generated in each setting.

The simulation study considered three settings with the following parameters held constant across data sets in the first setting: (1) number of clusters $N = 30$, chosen to ensure tractability of maximum likelihood estimation for the whole data, (2) number of experts $n = 147$, and (3) the fixed-effects β_j, α_j, sampled one time from a $N(0, 2)$ and $N(0, 1)$ respectively and then held constant in all data sets. Factors varying across the data sets were: the number of ratings per expert N_i and a set of 147 expert random-effects b_i. Conceptually, each generated data set represents a replication of the evaluation study in which a new set of experts rates the same clusters. Therefore, varying N_i and b_i from one data set to another resembles the use of different groups of experts in each study, sampled from the entire experts' population. The random expert specific effects b_i were sampled from $N(0, 10)$, the clusters evaluated by each expert were determined using the selection process $X_{ij}|b_i \sim$ Bernoulli (ρ_{ij}) with logit$(\rho_{ij}) = \alpha_j + b_i$ and the ratings $Y_{ij}|b_i$ were generated from a Bernoulli(π_{ij}) with logit$(\pi_{ij}) = \beta_j + b_i$.

Note that in the aforementioned setting the same expert specific random effects were used for the generation of the selection and rating processes and, consequently, the joint model

used in Section 5.1 to analyze this setting also satisfied this property, i.e., $a_i = b_i$. This is a special case of the general modeling framework introduced in Section 3, the so-called shared parameter model (SPM), in which the selection and rating processes shared a common random effect [8]. This simplified version was appealing to use in the simulation studies due to its reduced computational burden, since it only uses one set of random effects compared to two sets required for the general modeling framework.

To compare the performances of the combined and joint model in the most general and computationally demanding scenario presented in Section 3, a second setting was also considered on a smaller scale. In this second setting: (i) the total number of clusters was $N = 15$, (ii) the clusters evaluated by each expert were determined using the selection process $X_{ij}|a_i \sim$ Bernoulli (ρ_{ij}) with logit$(\rho_{ij}) = \alpha_j + a_i$, (iii) the rating process $Y_{ij}|b_i$ was generated from a Bernoulli(π_{ij}) with logit$(\pi_{ij}) = \beta_j + b_i$ and (iv) the random expert specific effects were generated as

$$\begin{bmatrix} b_i \\ a_i \end{bmatrix} \sim N \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10 & 4.95 \\ 4.95 & 5 \end{bmatrix} \right\}.$$

All the other parameters were like in the first setting. Following this data generating mechanism, the completely general version of the joint model, presented in in Section 3, was used in Section 5.1 to analyze these data.

In the two previous settings $P\big(Y_{ij} = y_{ij}|X_{ij} = 1, b_i\big) = P\big(Y_{ij} = y_{ij}|X_{ij} = 0, b_i\big)$ and, therefore, there is no selection bias in the rating process.

In order to evaluate the performance of both modelling approaches when there is a selection bias in the rating process, a third final setting was considered with the selection process generated as in the first setting, i.e., $X_{ij}|b_i \sim$ Bernoulli (ρ_{ij}) with logit$(\rho_{ij}) = \alpha_j + b_i$ and the rating probabilities generated as

$$\text{logit}\left[P(Y_{ij} = 1|X_{ij} = x_{ij}, b_i)\right] = \begin{cases} \beta_j + b_i & \text{if } x_{ij} = 1, \\ \beta_j + b_i - 0.223 & \text{if } x_{ij} = 0. \end{cases} \tag{13}$$

Here again all the other parameters were like in the first setting. Basically, (13) implies that, for every expert i, the odds of rating a cluster as 1 is 25% larger when the cluster is evaluated than when it is not. In this final setting the shared parameter version of the joint model was used to analyze the data in Section 5.2.

5.1 Results in absence of selection bias (settings 1 and 2)

Three analyses were carried out for each data set and the main results for the first setting, i.e., when the shared parameter model was used to generate the data, are summarized in Tables 1–4 (the results from the second setting are discussed at the end of this section). In

these tables, the column '*True*' always gives the true value of the corresponding parameter, the column '*Comb*' refers to results obtained from the SPM, the columns with 'J(·)' display results obtained from fitting the SPM using the selection probability derived from the logit in brackets and, finally, the column '*Naive*' presents the results obtained from fitting model (1) which does not account for the selection process.

Notice that model J$(\alpha_j + b_i)$ assumes that, for every expert, the selection probabilities vary with the clusters and the parameters governing the rating and selection processes are different. This model correctly described the data generating mechanism in the first setting of the simulations. In contrast, J$(\beta_j + b_i)$ also postulates different selection probabilities for the clusters but now the parametric space of the rating and selection processes are assumed to be equal. The last model J$(\alpha + b_i)$ presupposes equal selection probabilities for all the clusters conditional on the expert.

All models were fitted using maximum likelihood as implemented in the SAS procedure NLMIXED. The integrals in the likelihood function were approximated using the Gauss-Hermite quadrature method with 50 quadrature points. The Newton-Raphson algorithm was used to maximize the objective function and standard errors were derived from the corresponding Hessian matrix.

The results for the naive model clearly show that, even when there is no selection bias in the rating process, ignoring the selection process when constructing the likelihood for the estimation of the parameters can lead to extremely biased point estimates and confidence intervals with very low coverage probabilities. Interestingly, the correctly specified SPM J$(\alpha_j + b_i)$ also produced very biased point estimates for the parameters of interest and confidence intervals with low coverage. It is important to point out that the correctly specified SPM suffered heavily from lack-of-convergence problems and, therefore, numerical instability may be the reason behind its poor performance. The misspecified SPM J$(\alpha + b_i)$ produced very biased point estimates for some of the parameters of interest and the coverage probability of some confidence intervals were much smaller than the pre-specified 95%. However, in spite of these problems, this misspecified SPM clearly outperformed the correctly specified one in this scenario. Here again numerical stability seems to be the reason behind this awkward result. With 32 parameters, the misspecified model is computationally lighter than the correctly specified (61 parameters) and suffered from fewer convergence problems. Actually, in additional simulations with only 15 clusters (not shown) the correctly specified model unmistakably outperformed all the others, stressing the importance of numerical issues when analyzing this type of complex data.

The other misspecified SPM J$(\beta_j + b_i)$ had an extremely poor performance with relative biases larger than 1000% for some estimates and confidence intervals with null coverage probabilities. Incorrectly using the same set of parameters for the rating and selection model leads to estimates that try to describe both processes simultaneously and, hence, the rating modeled is

estimated with bias. Essentially, these results illustrate that misspecifying the selection model may have a huge negative impact on both point estimates and inferences.

Finally, the combined model always led to unbiased estimates of the parameters and confidence intervals with coverage probabilities close to the pre-specified 95%. The combined model was also the most stable numerically and its performance was the same when either 15 or 30 clusters were considered. In addition, while the average computation time for a single correctly specified SPM was 24 hours, for a single combined model the computational time was merely 40 minutes. This makes the combined model a competitive alternative to the joint model approach. It is important to point out that, while the use of random effects may help to avoid misspecification in the selection process, this level of generality also seems to produce less precise estimates in this setting, as illustrated by the larger standard errors in Table 1 and the wider confidence intervals in Table 3.

The results from the second setting, i.e, when the data were generated using different random effects for the selection and rating processes, as presented in Section 3, further vindicated the good performance of the combined model (results provided in the web supporting materials). Indeed, as in the previous setting, the combined model produced unbiased estimates for all parameters. In addition, the true joint model produced unbiased estimates for most parameters in this setting and its results were more precise than those of the combined model. Actually, the use of 15 instead of 30 clusters largely eliminated the numerical issues afflicting the correctly specified joint model and, hence, its performance was substantially improved. Analogous to the first setting, the misspecified joint model J$(\alpha + a_i)$ exhibited similar performance to the true joint model, while the joint model J$(\beta_j + a_i)$ performed the worst in terms of bias.

Another important difference between the joint and combined modelling approaches is the amount of information from the $\boldsymbol{X}_i$s they use. In fact, whereas the joint model uses all the information contained in $\boldsymbol{X}_i$ explicitly, the combined model only uses the information of the evaluated clusters and, hence, it only implicitly uses the information contained in $\boldsymbol{X}_i$. This may help to explain the more precise results produced by the joint model in the simulations with 15 clusters. However, this advantage comes with the caveat of additional modeling assumptions for the selection process and, as the simulations showed, violation of these assumptions may produce misleading results.

5.2 Results in the presence of selection bias (setting 3)

The results obtained when ratings are generated under model (13) were similar to those in Section 5.1 (results provided in the web supporting materials). The combined model produced unbiased estimates for both the cluster effects and probabilities of success. In some cases the SPM and naive model overestimated the probability of success. Despite these good results obtained for the combined model in presence of selection bias, one should be cautious when

generalizing these findings. The combined model assumes a very specific form of selection bias in the rating process, for instance as portrayed in (10), and this assumption may not always lead to satisfactory results. Obviously more theoretical research and simulation studies considering other mechanisms different from (13) will be necessary to assess the performance of the combined model in this scenario.

6 Case Study Analysis

The case study introduced in Section 2 was analyzed by Alonso *et al.* [6] using the naive and joint model approaches. In the present work the combined model presented in Section 4 was also fitted to these data. The shared random effects assumption, underlying the SPM, implies that experts who rate more clusters will tend to give more positive recommendations as well. This testable assumption was not supported by data of the case study and, therefore, the SPM was not used in the analysis.

Owing to the high dimensionality problem discussed in Milanzi *et al.* [5], maximum likelihood became infeasible and the analysis of the case study was performed using the procedure proposed by the same authors, which basically involves 4 steps. First, the dataset is split into mutually exclusive and exhaustive subsets of clusters and their corresponding ratings. This ensures that each subset has the required information to estimate the effect of clusters contained therein. In the second step, the models as described in Sections 3 and 4 are fitted in each subset using maximum likelihood to obtain the cluster-effects estimates. Collecting the estimates from all subsets produces one set of estimates for all clusters. The third step involves permutations of the dataset (W permutations) and repeating steps 1 and 2 for each permutation. In the final step, estimates for each cluster obtained from the different permutations are pooled together to obtain the final estimate. In spite of its relevance for the present work, a lengthy discussion of this procedure would go beyond the scope of the manuscript and we refer the interested reader to Milanzi *et al.* [5] for more details.

The main findings are displayed in Table 5 where the clusters are ordered according to the ranks obtained from the combined model. Clearly, many clusters are ranked consistently high by the three approaches (for instance clusters 295061, 296535, 84163, etcetera) and, therefore, should be given priority. Along these lines, in a sensitivity analysis context, a cut-off point can be agreed upon in terms of probability or rank as illustrated in Figure 2. The figure shows the top 1000 clusters according to the combined model where, for illustration purposes, cut-off points of 500 and 0.5 were set for the rank and probability of success, respectively. Using these criteria, researchers can decide to give priority to clusters that are ranked in the top 500 by all methods. Alternatively, the decision making could be based on the probabilities of success directly and clusters whose estimated probability is 0.5 and above for all methods could be selected for further scrutiny. Notice that both approaches led to a substantial reduction of

the total number of clusters that would have to be analyzed in future studies. Scientific and practical considerations will certainly play a role when determining the cut-off points in a real decision making process.

On the other hand, for other clusters the three approaches led to strikingly different results. Indeed, the fourth best cluster according to the naive approach (69850) received ranks 182 and 38 from the joint and combined models respectively. Moreover, compound 265222 ranked first and third by the naive and joint models, respectively, was not among the top ten clusters according to the combined model. Arguably, these disparate results for some clusters confirm the utility of basing the decision making process not in a unique modeling approach but several ones, i.e., the utility of a sensitivity analysis.

Several strategies could be implemented when further exploring these clusters for which the statistical evidence was inconclusive. For instance, one could compute the average probability of success over the different approaches and base the selection on this average probability. On the other hand, given the results of the simulations one could argue that, since the combined model seems to produced unbiased estimates in most circumstances, it should be the core of the decision making process. Whatever strategy is finally adopted a careful discussion with the experts in the field would always be advisable. Eventually, weighting together the quantitative results obtained from the statistical analysis and more field specific knowledge may help to make an optimal and thoughtful choice.

Interestingly, although the estimates from the combined model were less precise in the simulations, in the case study it produced estimates as precise as the ones obtained in the other modelling frameworks. Probably, the larger amount of information available in the case study and the greater numerical stability of the combined model made the previously observed difference in precision negligible.

7 Conclusion

We have introduced an alternative approach for the evaluation of clusters of chemical compounds, based on the so-called combined model. The model accounts for the selection process using a new set of random effects. Simulation results clearly showed that, unlike the naive and joint model approaches, the combined model seems to produce nearly unbiased estimates in most settings and exhibited important numerical advantages.

The novel idea of implicitly accounting for the presence of selection bias in the rating process by adding latent traits, makes the combined model a valuable addition to the literature covering selection bias methods for repeated measures. More often than not, the selection process is a nuisance, hence, estimation of parameters specific to the selection process can be viewed as a waste of degrees of freedom. By treating the selection process as a latent trait nuisance,

the combined model minimizes this wastage, avoids the need of using explicit models to describe the selection process and, as the simulation study showed, significantly increases numerical performance. For instance, in the simulations the SPM needed to estimate 31 parameters specific to the selection process while the combined model only had to estimate one. Additionally, the impact of misspecification (selection model, for joint model and distribution of random effects for the selection process, for the combined model) is more likely to be felt in the former than in the latter as also observed in simulation results.

Given the robustness and numerical stability exhibited by the combined model we believe that, even when selection bias is not suspected and the factors that drive the selection process are known and available, one may still want to use the combined model as a sensitivity tool for the analysis.

It is worth mentioning that by only using information from the evaluated clusters, the way in which the combined model accounts for the presence of a selection bias in the rating process falls under the category of conditioning on a common effect as explained in [17], where the selection process is the common effect. This has been well-illustrated in Section 4. Hernán, Hernández-Diaz, and Robins [17] further suggest methods to tackle such bias which include semi-parametric methods, such as inverse probability weighting and their doubly-robust extensions. There are similarities between the combined model approach and how selection bias is dealt with in the causal inference literature. In fact, as shown in the web supporting materials, in the combined model approach the probability of the counterfactual event $\{Y_{ij} = 1|X_{ij} = 0, b_i\}$ is implicitly embedded in the structure of the model. The use of counterfactuals lies at the core of many causal inference methods and the connection between these procedures and the combined model are worth exploring. In addition, Bayesian methods are particularly suited to handle situations where a large number of sources of uncertainty need to be taken into account and their computational flexibility can allow the use of non conjugate distributions for the latent selection traits in the combined model. Even though the implementation of a Bayesian approach clearly surpasses the scope of this work, exploring this alternative is certainly worth pursuing.

It is important to point out that our modeling approach can be given further uses not discussed here. For example, in addition to rating compounds, also the rater could be rated. Apart from intrinsic individual differences, also practice and experience can be brought out, if present. Indeed, it is conceivable that there is a learning curve on the one hand, but also fatigue from 'over-rating' on the other.

Obviously, more theoretical developments, simulations, and the analysis of case studies will be needed to fully understand the potential and limitations of the approaches studied in this paper.

Acknowledgment

Elasma Milanzi and Geert Molenberghs gratefully acknowledge support from IAP research Network P7/06 of the Belgian Government (Belgian Science Policy). The authors are grateful to Johnson & Johnson for the kind permission to use their data, and David Amwonya for performing the extra simulations. For the computations, simulations and data processing, we used the infrastructure of the VSC — Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government — department EWI.

References

1. Alonso A, Molenberghs G. Surrogate endpoints: Hopes and perils. *Pharmacoeconomics and Outcomes Research* 2008; **3**: 255-259, Doi: 10.1586/14737167.8.3.255.

2. Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. *Lancet* 2007; **369**: 1883-1889.

3. Hack MD, Rassokhin DN, Buyck C, Seierstad M, Skalkin A, ten Holte P, Jones TK, Mirzadegan T, Agrafiotis DK. Library enhancement through the wisdom of crowds. *Journal of Chemical Information and Modeling* 2011; **51**: 3275-3286.

4. Milanzi, E. Flexible modeling for hierarchical data, data with random sample sizes and selection bias, with applications in pharmaceutical research Web. Sep. 2013. [https://ibiostat.be/publications/phd/elasmamilanzi.pdf].

5. Milanzi E, Alonso A, Buyck C, Molenberghs G, Bijnens L. A permutational-splitting sample procedure to quantify expert opinion on chemical cluster using high-dimensional data. *Annals of Applied Statistics* 2014; **00**: 00–00.

6. Alonso A, Milanzi E, Molenberghs G, Buyck C, and Bijnens L. Impact of selection bias on the qualitative assessment of biomolecular cluster. *Submitted for publication* 2013.

7. Creemers A, Hens N, Aert M, Molenbergh G, Verbeke G, Kenward MG. Generalized shared-parameter models and missingness at random. *Statistical Modeling*, 2011; **11**: 279-311.

8. Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics* 1995; **51**: 151-168.

9. Rizopoulos D, Verbeke G, Molenberghs G. Shared parameter models under random effects misspecification. *Biometrika* 2008; **95**: 63-74.

10. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data, With Applications in R*. Chapman and Hall/CRC: Boca Raton, 2012.

11. Vonesh EF, Green T, Schluchter MD. Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine* 2006; **25**: 143-163.

12. Genelletti S, Mason A, Best N. Adjusting for selection effects in epidemiological studies; Why sensitivity analysis is the only "solution". *Commentary in Epidemiology* 2011; **22**: 36-39.

13. Booth JG, Casella G, Friedl H, Hobert JP. Negative binomial loglinear mixed models. *Statistical Modelling* 2003; **3**: 179-181.

14. Molenberghs G, Verbeke G, Demétrio C, Vieira A. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science* 2010; **25**: 325-347.

15. Agrafiotis DK, Alex S, Dai H, Derkinderen A, Farnum M, Gates P, Izrailev S, Jaeger EP, Konstant P, Leung A, Lobanov VS, Marichal P, Martin D, Rassokhin DN, Shemanarev M, Skalkin A, Stong J, Tabruyn T, Vermeiren M, Wan J, Xu XY, Yao X. Advanced Biological and Chemical Discovery (ABCD): Centralizing discovery knowledge in an inherently decentralized world. *Journal of Chemical Information Modeling* 2007; **47**: 1999-2014.

16. Frederic P, Lad F. Two moments of the logitnormal distribution. *Communications in Statistics, Simulation and Computation* 2008; **37**: 1263-1269.

17. Hernán MA, Hernández-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**: 615–625.

Table 1: *Mean of the point estimates and (standard errors) from setting 1. True: true cluster-effect value, Comb: combined model, $J(\alpha_j + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \alpha_j + b_i$, $J(\beta_j + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \beta_j + b_i$, $J(\alpha + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \alpha + b_i$, Naive: Naive model.*

Parameter	True	Comb		$J(\alpha_j + b_i)$		$J(\beta_j + b_i)$		$J(\alpha + b_i)$		Naive	
β_1	3.60	3.60	(0.87)	4.17	(0.72)	0.20	(0.33)	3.66	(0.84)	4.71	(0.84)
β_2	-1.98	-1.98	(0.62)	-0.36	(0.22)	-0.37	(0.27)	-1.97	(0.37)	1.33	(0.37)
β_3	4.33	4.33	(0.85)	3.89	(0.71)	2.04	(0.34)	4.15	(0.71)	5.11	(0.79)
β_4	0.58	0.58	(0.65)	1.80	(0.39)	-0.47	(0.32)	0.54	(0.50)	1.47	(0.53)
β_5	0.11	0.11	(0.62)	1.48	(0.35)	-0.54	(0.31)	0.10	(0.46)	0.96	(0.48)
β_6	-0.53	-0.52	(0.63)	1.13	(0.31)	-0.77	(0.31)	-0.46	(0.45)	0.37	(0.46)
β_7	1.70	1.70	(0.91)	2.94	(0.72)	-1.00	(0.33)	1.50	(0.71)	2.71	(0.80)
β_8	-0.10	-0.10	(0.57)	1.17	(0.28)	-0.14	(0.30)	-0.12	(0.41)	0.67	(0.42)
β_9	1.51	1.51	(0.76)	2.72	(0.52)	-0.48	(0.32)	1.54	(0.64)	2.55	(0.70)
β_{10}	1.29	1.30	(0.65)	2.16	(0.40)	0.09	(0.32)	1.24	(0.52)	2.12	(0.55)
β_{11}	0.88	0.88	(0.87)	2.53	(0.60)	-1.49	(0.33)	0.89	(0.73)	1.86	(0.74)
β_{12}	-3.52	-3.52	(0.82)	-1.01	(0.26)	-1.26	(0.26)	-3.41	(0.42)	2.73	(0.42)
β_{13}	0.60	0.60	(0.58)	1.57	(0.33)	-0.09	(0.31)	0.48	(0.44)	1.43	(0.47)
β_{14}	1.89	1.89	(0.93)	3.65	(0.72)	-1.01	(0.33)	1.84	(0.79)	2.88	(0.81)
β_{15}	0.68	0.69	(0.57)	1.52	(0.31)	0.15	(0.31)	0.60	(0.43)	1.49	(0.46)
β_{16}	0.05	0.05	(0.56)	1.11	(0.27)	-0.12	(0.30)	-0.13	(0.40)	0.77	(0.42)
β_{17}	0.11	0.12	(0.48)	0.63	(0.21)	0.92	(0.28)	0.02	(0.36)	0.72	(0.36)
β_{18}	-4.01	-4.01	(0.87)	-1.39	(0.27)	-1.12	(0.24)	-3.95	(0.43)	3.27	(0.43)
β_{19}	0.27	0.27	(0.55)	1.18	(0.28)	0.10	(0.30)	0.11	(0.41)	1.00	(0.43)
β_{20}	-1.79	-1.79	(0.59)	-0.39	(0.21)	-0.14	(0.26)	-1.82	(0.36)	1.14	(0.36)
β_{21}	1.03	1.03	(0.57)	1.58	(0.31)	0.45	(0.31)	0.90	(0.43)	1.70	(0.46)
β_{22}	0.03	0.03	(0.58)	1.30	(0.32)	-0.39	(0.31)	0.02	(0.44)	0.81	(0.45)
β_{23}	-0.05	-0.05	(0.59)	1.30	(0.32)	-0.37	(0.31)	-0.06	(0.43)	0.79	(0.45)
β_{24}	-0.95	-0.95	(0.54)	0.31	(0.22)	-0.06	(0.28)	-0.98	(0.37)	0.30	(0.37)
β_{25}	0.10	0.10	(0.63)	1.43	(0.34)	-0.63	(0.31)	0.08	(0.46)	0.93	(0.49)
β_{26}	0.87	0.88	(0.58)	1.69	(0.33)	0.19	(0.31)	0.80	(0.45)	1.61	(0.46)
β_{27}	2.13	2.13	(1.01)	3.31	(0.72)	-1.41	(0.35)	2.36	(0.83)	3.28	(0.89)
β_{28}	-3.03	-3.03	(0.73)	-0.88	(0.24)	-0.75	(0.25)	-3.05	(0.39)	2.37	(0.39)
β_{29}	-0.34	-0.33	(0.54)	0.79	(0.25)	0.07	(0.29)	-0.36	(0.38)	0.42	(0.39)
β_{30}	2.06	2.06	(0.74)	2.64	(0.51)	0.26	(0.32)	1.78	(0.59)	2.89	(0.65)
σ^2	10.00	10.14	(5.20)	10.05	(1.54)	7.61	(1.09)	7.99	(1.15)	5.93	(1.19)
κ		0.0007	(0.0814)								

Table 2: *Mean relative bias for setting 1. Comb: combined model, $J(\alpha_j + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \alpha_j + b_i$, $J(\beta_j + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \beta_j + b_i$, $J(\alpha + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \alpha + b_i$, Naive: Naive model.*

Parameter	Comb	$J(\alpha_j + b_i)$	$J(\beta_j + b_i)$	$J(\alpha + b_i)$	Naive
β_1	0.00	0.16	0.94	0.02	0.31
β_2	0.00	0.82	0.81	0.00	0.33
β_3	0.00	0.10	0.53	0.04	0.18
β_4	0.00	2.10	1.80	0.07	1.53
β_5	0.02	12.82	6.09	0.06	7.96
β_6	0.01	3.15	0.46	0.13	1.70
β_7	0.00	0.73	1.59	0.12	0.59
β_8	0.03	12.59	0.36	0.21	7.67
β_9	0.00	0.81	1.32	0.02	0.69
β_{10}	0.00	0.67	0.93	0.04	0.64
β_{11}	0.00	1.88	2.70	0.02	1.12
β_{12}	0.00	0.71	0.64	0.03	0.22
β_{13}	0.00	1.60	1.14	0.19	1.38
β_{14}	0.00	0.93	1.53	0.03	0.52
β_{15}	0.00	1.22	0.78	0.13	1.18
β_{16}	0.06	20.75	3.37	3.51	14.09
β_{17}	0.03	4.54	7.03	0.81	5.34
β_{18}	0.00	0.65	0.72	0.02	0.18
β_{19}	0.01	3.43	0.61	0.59	2.77
β_{20}	0.00	0.78	0.92	0.02	0.36
β_{21}	0.00	0.53	0.56	0.13	0.65
β_{22}	0.10	48.42	15.80	0.06	29.59
β_{23}	0.06	28.09	6.75	0.33	17.52
β_{24}	0.00	1.33	0.93	0.02	0.69
β_{25}	0.02	12.96	7.16	0.22	8.08
β_{26}	0.00	0.93	0.78	0.08	0.85
β_{27}	0.00	0.56	1.66	0.11	0.54
β_{28}	0.00	0.71	0.75	0.01	0.22
β_{29}	0.01	3.34	1.20	0.09	2.25
β_{30}	0.00	0.28	0.87	0.14	0.40
σ^2	0.01	0.01	0.24	0.20	0.41

Table 3: *Mean confidence interval coverage and length for setting 1. Comb: combined model, $J(\alpha_j + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \alpha_j + b_i$, $J(\beta_j + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \beta_j + b_i$, $J(\alpha + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \alpha + b_i$, Naive: Naive model.*

Parameter	Comb		$J(\alpha_j + b_i)$		$J(\beta_j + b_i)$		$J(\alpha + b_i)$		Naive	
β_1	0.98	(12.87)	0.78	(2.77)	0.00	(1.30)	0.66	(3.09)	0.81	(3.58)
β_2	0.94	(5.30)	0.04	(0.89)	0.00	(1.04)	0.94	(1.45)	0.60	(1.45)
β_3	0.98	(8.47)	0.70	(2.88)	0.00	(1.32)	0.83	(3.02)	0.94	(3.36)
β_4	0.94	(4.32)	0.04	(1.75)	0.05	(1.25)	0.94	(2.03)	0.63	(2.17)
β_5	0.93	(3.64)	0.04	(1.35)	0.47	(1.23)	0.95	(1.84)	0.61	(1.94)
β_6	0.94	(4.04)	0.04	(1.29)	0.90	(1.22)	0.95	(1.75)	0.51	(1.82)
β_7	0.96	(6.67)	0.70	(2.79)	0.00	(1.31)	0.86	(2.90)	0.86	(3.29)
β_8	0.93	(3.73)	0.04	(1.16)	0.97	(1.19)	0.97	(1.62)	0.60	(1.68)
β_9	0.96	(5.42)	0.30	(2.26)	0.00	(1.28)	0.92	(2.60)	0.75	(2.84)
β_{10}	0.97	(3.76)	0.30	(1.69)	0.00	(1.25)	0.94	(2.11)	0.71	(2.25)
β_{11}	0.97	(6.31)	0.09	(2.71)	0.00	(1.32)	0.89	(2.95)	0.83	(3.15)
β_{12}	0.94	(7.68)	0.04	(1.09)	0.00	(1.02)	0.96	(1.65)	0.54	(1.66)
β_{13}	0.95	(3.59)	0.13	(1.33)	0.44	(1.22)	0.93	(1.77)	0.61	(1.88)
β_{14}	0.98	(6.58)	0.57	(2.58)	0.00	(1.32)	0.86	(7.36)	0.86	(3.40)
β_{15}	0.94	(3.17)	0.13	(1.25)	0.51	(1.21)	0.94	(1.72)	0.60	(1.82)
β_{16}	0.95	(3.39)	0.00	(1.10)	0.94	(1.18)	0.93	(1.59)	0.61	(1.67)
β_{17}	0.94	(2.71)	0.17	(0.87)	0.17	(1.11)	0.95	(1.40)	0.62	(1.41)
β_{18}	0.95	(8.19)	0.09	(1.15)	0.00	(0.97)	0.94	(1.70)	0.58	(1.69)
β_{19}	0.94	(3.23)	0.04	(1.12)	0.91	(1.19)	0.94	(1.61)	0.62	(1.68)
β_{20}	0.95	(4.97)	0.00	(0.87)	0.00	(1.03)	0.97	(1.42)	0.58	(1.41)
β_{21}	0.95	(2.80)	0.48	(1.25)	0.53	(1.21)	0.90	(1.72)	0.70	(1.82)
β_{22}	0.94	(3.79)	0.00	(1.28)	0.73	(1.21)	0.95	(1.73)	0.60	(1.79)
β_{23}	0.95	(3.77)	0.04	(1.27)	0.85	(1.21)	0.98	(1.70)	0.54	(1.79)
β_{24}	0.93	(4.18)	0.00	(0.90)	0.08	(1.10)	0.96	(1.45)	0.58	(1.46)
β_{25}	0.95	(3.95)	0.04	(1.38)	0.38	(1.24)	0.97	(1.82)	0.60	(1.94)
β_{26}	0.96	(3.40)	0.17	(1.37)	0.43	(1.22)	0.97	(1.76)	0.64	(1.87)
β_{27}	0.95	(10.32)	0.57	(2.19)	0.00	(1.36)	0.65	(4.77)	0.68	(3.44)
β_{28}	0.97	(6.67)	0.04	(0.98)	0.00	(0.99)	0.96	(1.55)	0.60	(1.54)
β_{29}	0.95	(3.64)	0.00	(1.00)	0.77	(1.15)	0.94	(1.51)	0.52	(1.55)
β_{30}	0.97	(4.56)	0.83	(2.06)	0.00	(1.26)	0.94	(5.59)	0.81	(2.69)

Table 4: *Mean probability estimates and relative bias for setting 1. True: true value, Comb: combined model, $J(\alpha_j + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \alpha_j + b_i$, $J(\beta_j + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \beta_j + b_i$, $J(\alpha + b_i)$: joint model with logit$[P(X_{ij} = 1|b_i)] = \alpha + b_i$, Naive: Naive model.*

Parameter	True	Comb		$J(\alpha_j + b_i)$		$J(\beta_j + b_i)$		$J(\alpha + b_i)$		Naive	
β_3	0.88	0.88	(0.01)	0.86	(0.02)	0.73	(0.16)	0.89	(0.02)	0.96	(0.09)
β_1	0.83	0.84	(0.01)	0.87	(0.05)	0.52	(0.37)	0.86	(0.04)	0.94	(0.13)
β_{27}	0.71	0.72	(0.01)	0.86	(0.21)	0.34	(0.53)	0.76	(0.08)	0.86	(0.21)
β_{30}	0.70	0.71	(0.01)	0.76	(0.07)	0.53	(0.25)	0.70	(0.00)	0.84	(0.19)
β_{14}	0.69	0.70	(0.01)	0.84	(0.22)	0.38	(0.45)	0.71	(0.03)	0.83	(0.21)
β_7	0.67	0.68	(0.02)	0.79	(0.18)	0.38	(0.43)	0.68	(0.01)	0.82	(0.22)
β_9	0.65	0.66	(0.02)	0.77	(0.19)	0.44	(0.32)	0.68	(0.05)	0.81	(0.24)
β_{10}	0.63	0.64	(0.02)	0.72	(0.15)	0.51	(0.19)	0.65	(0.03)	0.76	(0.21)
β_{21}	0.60	0.61	(0.02)	0.67	(0.12)	0.56	(0.08)	0.61	(0.01)	0.72	(0.20)
β_{11}	0.58	0.60	(0.02)	0.77	(0.31)	0.33	(0.44)	0.60	(0.03)	0.73	(0.25)
β_{26}	0.58	0.59	(0.02)	0.68	(0.17)	0.52	(0.10)	0.60	(0.03)	0.71	(0.22)
β_{15}	0.56	0.57	(0.02)	0.67	(0.18)	0.52	(0.08)	0.57	(0.02)	0.69	(0.22)
β_{13}	0.55	0.57	(0.02)	0.67	(0.21)	0.49	(0.12)	0.56	(0.01)	0.68	(0.23)
β_4	0.55	0.56	(0.02)	0.69	(0.24)	0.44	(0.20)	0.56	(0.02)	0.69	(0.25)
β_{19}	0.52	0.53	(0.02)	0.63	(0.22)	0.51	(0.01)	0.51	(0.01)	0.63	(0.22)
β_{17}	0.50	0.51	(0.03)	0.57	(0.14)	0.61	(0.22)	0.50	(0.00)	0.59	(0.19)
β_5	0.50	0.51	(0.02)	0.67	(0.33)	0.44	(0.13)	0.51	(0.03)	0.63	(0.25)
β_{25}	0.50	0.51	(0.03)	0.65	(0.30)	0.43	(0.14)	0.51	(0.02)	0.62	(0.25)
β_{16}	0.49	0.51	(0.03)	0.62	(0.25)	0.49	(0.02)	0.48	(0.02)	0.60	(0.23)
β_{22}	0.49	0.50	(0.03)	0.64	(0.29)	0.45	(0.07)	0.50	(0.02)	0.61	(0.23)
β_{23}	0.48	0.49	(0.03)	0.65	(0.35)	0.46	(0.06)	0.49	(0.02)	0.60	(0.25)
β_8	0.48	0.49	(0.03)	0.63	(0.31)	0.48	(0.02)	0.49	(0.02)	0.59	(0.24)
β_{29}	0.45	0.46	(0.03)	0.58	(0.29)	0.51	(0.13)	0.46	(0.01)	0.56	(0.23)
β_6	0.43	0.44	(0.03)	0.62	(0.45)	0.41	(0.05)	0.45	(0.04)	0.55	(0.27)
β_{24}	0.38	0.40	(0.03)	0.53	(0.39)	0.49	(0.28)	0.38	(0.00)	0.46	(0.19)
β_{20}	0.30	0.31	(0.04)	0.46	(0.53)	0.48	(0.61)	0.29	(0.02)	0.35	(0.17)
β_2	0.28	0.29	(0.04)	0.46	(0.64)	0.46	(0.61)	0.28	(0.02)	0.33	(0.16)
β_{28}	0.19	0.20	(0.05)	0.40	(1.05)	0.41	(1.12)	0.18	(0.06)	0.21	(0.11)
β_{12}	0.16	0.17	(0.06)	0.39	(1.49)	0.35	(1.21)	0.16	(0.01)	0.18	(0.15)
β_{18}	0.13	0.14	(0.06)	0.35	(1.75)	0.37	(1.85)	0.12	(0.06)	0.14	(0.08)

Table 5: *Estimated parameters ($\hat{\beta}$), and standard errors for the top 20 clusters (according to the combined model approach) from the case study. The models fitted are: Combined model (Combined), mixed logistic regression (Naive), and joint model with selection probability given by logit$[P(X_{ij} = 1|a_i)] = \alpha_j + a_i$ [$J(\alpha_j + a_i)$]. The column CID gives the cluster id.*

	Naive		$J(\alpha_j + a_i)$		Combined	
CID	$\hat{\beta}$	Std.Error	$\hat{\beta}$	Std.Error	$\hat{\beta}$	Std.Error
295061	3.83	1.38	2.61	1.01	1.69	0.89
296535	2.77	1.35	2.37	1.75	1.58	1.02
333529	1.68	2.72	1.12	3.15	1.51	1.03
313914	2.18	1.58	2.06	1.68	1.47	1.33
356662	1.46	1.45	0.77	1.73	1.41	1.12
84163	5.24	2.42	1.83	3.05	1.34	1.40
296427	1.63	1.94	1.17	2.02	1.35	1.34
263047	1.70	1.26	0.67	1.19	1.32	1.00
315928	2.04	1.52	1.47	1.41	1.26	0.92
150535	1.12	1.48	0.68	1.64	1.23	1.05
292579	1.85	1.95	1.50	2.30	1.23	1.07
465585	-0.10	1.26	-0.17	1.39	1.20	1.23
178994	1.85	1.20	1.57	1.12	1.14	0.89
338571	1.08	1.26	0.84	1.32	1.12	1.14
296560	1.89	0.99	1.86	1.07	1.09	1.02
7608	1.38	0.95	0.83	1.20	1.09	1.03
178828	1.55	1.06	1.42	1.22	1.08	1.06
483662	1.08	1.31	0.10	1.66	1.09	1.23
383873	1.20	1.35	1.01	1.45	1.08	1.05
292805	1.47	1.24	1.20	1.21	1.06	1.09
σ_b^2	20.02	29.39	18.61	25.81	6.64	1.59
σ_a^2			5.4	1.7		
κ					0.0001	0.05

Table 6: *Estimated rank, probability of success ($\hat{P}$), and 95% confidence interval for the top 20 clusters (according to the combined model approach) from the case study. The models fitted are: Combined model (Combined), mixed logistic regression (Naive), and joint model with selection probability given by logit$[P(X_{ij} = 1|a_i)] = \alpha_j + a_i$ [$J(\alpha_j + a_i)$]. The column CID gives the cluster id.*

	Naive				$J(\alpha_j + a_i)$				Combined			
CID	R	$\hat{P}$	95% CI		R	$\hat{P}$	95% CI		R	$\hat{P}$	95% CI	
295061	2	0.92	0.66	0.98	4	0.71	0.48	0.87	1	0.71	0.48	0.86
296535	10	0.71	0.48	0.87	5	0.70	0.38	0.91	2	0.70	0.44	0.87
333529	21	0.63	0.31	0.87	26	0.59	0.22	0.99	3	0.69	0.43	0.87
313914	11	0.70	0.40	0.89	7	0.68	0.28	0.91	4	0.68	0.35	0.89
356662	23	0.63	0.31	0.85	40	0.57	0.23	0.86	5	0.67	0.39	0.87
84163	5	0.77	0.41	0.97	9	0.65	0.21	0.97	6	0.67	0.32	0.89
296427	24	0.62	0.27	0.87	22	0.60	0.19	0.90	7	0.67	0.33	0.89
263047	27	0.62	0.35	0.83	38	0.57	0.29	0.81	8	0.67	0.41	0.85
315928	9	0.72	0.37	0.92	14	0.62	0.28	0.83	9	0.66	0.42	0.83
150535	43	0.60	0.31	0.83	51	0.56	0.23	0.84	10	0.65	0.39	0.85
292579	19	0.64	0.24	0.90	13	0.63	0.21	0.89	11	0.65	0.39	0.85
465585	98	0.55	0.28	0.79	281	0.48	0.20	0.78	12	0.65	0.35	0.87
178994	13	0.68	0.45	0.84	11	0.64	0.34	0.84	13	0.65	0.43	0.82
338571	36	0.60	0.34	0.82	33	0.57	0.28	0.82	14	0.64	0.36	0.85
296560	14	0.66	0.43	0.83	8	0.66	0.39	0.85	15	0.64	0.38	0.83
7608	30	0.61	0.39	0.80	41	0.57	0.31	0.79	16	0.64	0.38	0.83
178828	25	0.62	0.38	0.81	17	0.61	0.32	0.84	17	0.64	0.38	0.83
483662	78	0.56	0.30	0.81	153	0.51	0.24	0.80	18	0.64	0.35	0.85
383873	33	0.61	0.32	0.83	27	0.59	0.28	0.84	19	0.64	0.38	0.83
292805	17	0.65	0.38	0.84	19	0.61	0.29	0.84	20	0.64	0.37	0.84

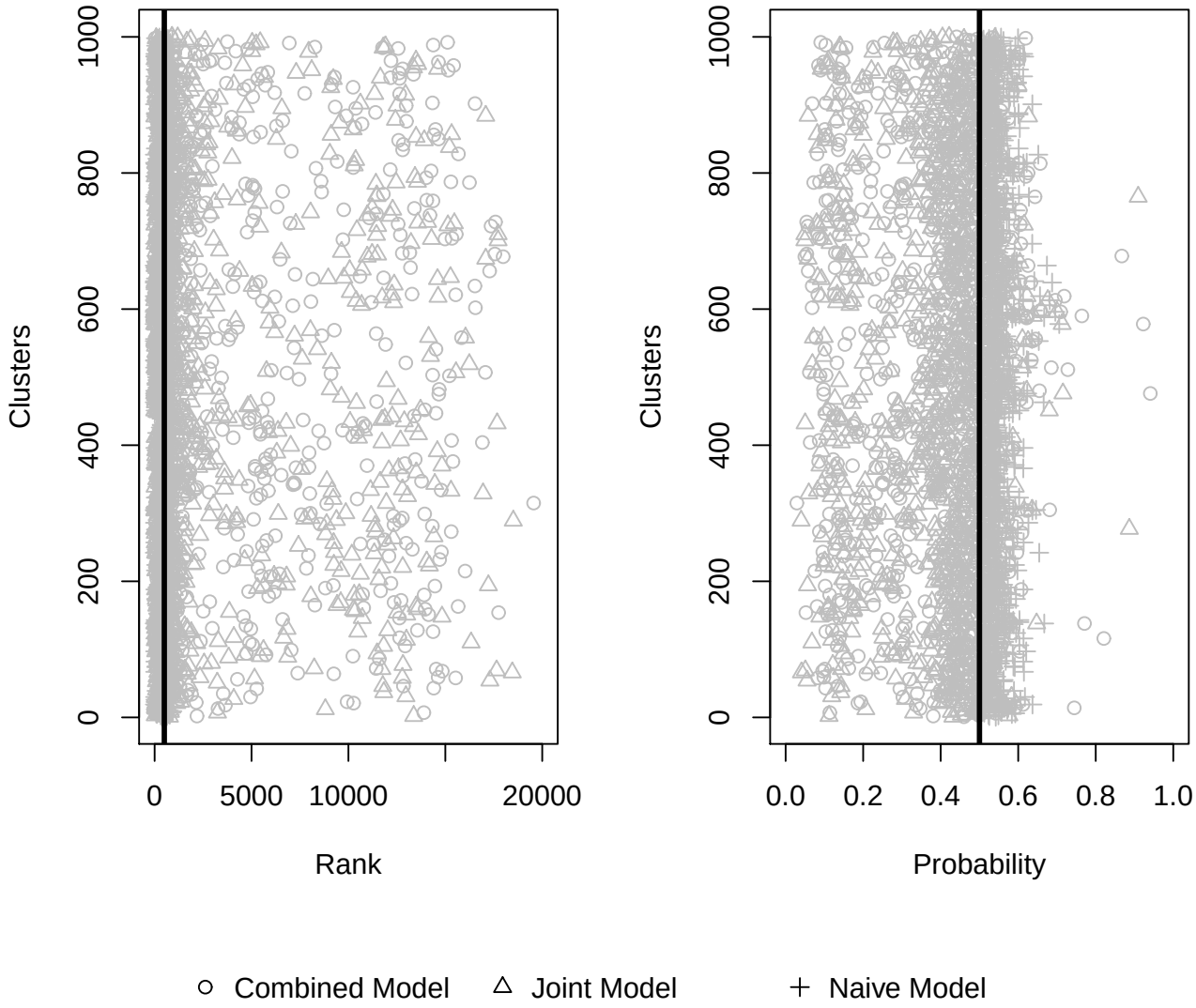

Figure 2: *A plot for the top 1000 clusters (according to the combined model). On the left: Plot of the ranks given to the clusters by the three models (combined, joint and naive) with a cut-off point at rank=500. On the right: Plot of probabilities of success given to the 1000 clusters by the three models, with a cut-off point at probability=0.5.*