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

MILANZI, Elasma; ALONSO ABAD, Ariel; MOLENBERGHS, Geert; Buyck, Christophe & BIJNENS, Luc (2015) A new modeling approach for quantifying expert opinion in the drug discovery process. In: STATISTICS IN MEDICINE, 34 (9), p. 1590-1604.

DOI: 10.1002/sim.6459
Handle: http://hdl.handle.net/1942/18650

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

Ariel Alonso[1] **Elasma Milanzi**[2] **Geert Molenberghs**[2,3]
Christophe Buyck[4] **Luc Bijnens**[4]

[1] *Department of Methodology and Statistics. Maastricht University. The Netherlands*

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

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

[4] *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 in practice, the process by which the clusters are assigned to the experts for evaluation, the so-called selection process, and the qualitative ratings given by them (chosen/not chosen) need to be jointly modeled in order 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 the scenarios considered. In contrast, the estimates from the joint model were severely biased under some misspecifications 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, due to its greater numerical stability, the combined model may outperform the joint model even when the latter is correctly specified.

Keywords: Selection bias, Combined model, 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 fact, 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 selected 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 absence of selection bias, one often needs to jointly model the rating and selection processes to get valid estimates. Ideally, one would like to know all the factors influencing the selection process before hand. However, in practice, such information is seldom available and making assumptions on the selection process is then almost 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 described, 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 via simulation. 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 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 the

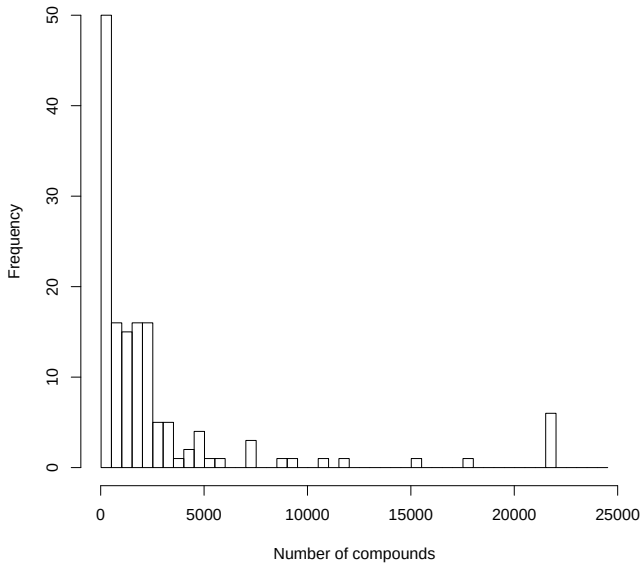

Figure 1: *Histogram for the number of clusters rated by the experts. The height of a bar indicates the number of experts whose number of rated cluster fall within the range given by the width of the bar.*

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 reassuming 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 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.,

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

where $\phi(b|0, \sigma_b^2)$ 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 [4] and Milanzi *et al.* [5] developed an algorithm that allows to handle these issues with a very small loss of efficiency and in the present work the dimensionality problem will not be discussed further.

Alonso *et al.* [6] pointed out that model (1) actually 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. Furthermore, they assumed that the vectors of expert-specific random effects $(a_i, b_i)^T$ are independent and 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 $P\big(Y_{ij} = y_{ij}|X_{ij} = 1, b_i\big) \neq P\big(Y_{ij} = y_{ij}|X_{ij} = 0, b_i\big)$ and showed that absence of selection bias 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

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

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

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

Clearly, there is information about how the experts rated the clusters they evaluated and, therefore, $P\left(Y_{ij} = 1 | X_{ij} = 1, b_i\right)$ can be estimated from the data using model (1). Furthermore, there is also information about which clusters every expert evaluated and this information could be used to estimate $P\left(X_{ij} = 1 | a_i\right)$. 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 $P\left(Y_{ij} = 1 | X_{ij} = 0, b_i\right)$ 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). Like Alonso *et al.* [6] 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

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

where

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

and $P(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta}) = \prod_j^{N_i} P(Y_{1ij} = y_{1ij}|b_i, \beta_j)$, $P(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\alpha}) = \prod_j^{N} P(X_{ij} = x_{ij}|a_i, \alpha_j)$ [4, 6].

Using the maximum likelihood estimators $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\boldsymbol{\alpha}}_n$, $\widehat{\sigma}^2_{bn}$ obtained from (6), one can estimate the probabilities of success by substituting $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\sigma}^2_{bn}$ into (2). Note, however, that to estimate $\boldsymbol{\beta}$, σ^2_b, 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 $\phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}) = \phi(a_i|0, \sigma^2_a)\phi(b_i|0, \sigma^2_b)$ and have a disjoint parametric space. In fact, under these assumptions (7) simplifies to

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

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

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

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 fully ignored. This setting will result, for instance, if a fully random allocation of the cluster to raters is implemented, so that the raters have no influence whatsoever on the selection process. However, if the raters can influence the selection process then a selection model will need to be incorporated into the analysis to guaranty valid results, even if selection bias is not present.

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 $\left(y_{ij}, \theta_{ij}, b_i\right)^T$ 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 rater effect b_i. It is important to point out that, although θ_{ij} and b_i are independent, 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

$$\begin{aligned} Y_{ij} | b_i, \theta_{ij} \sim \text{Bernoulli}(\theta_{ij}\pi_{ij}), \quad \pi_{ij} =& \frac{\exp(\beta_j + b_i)}{1 + \exp(\beta_j + b_i)}, \\ \theta_{ij} \sim \text{Beta}\left(\lambda, \tau\right), \qquad\qquad b_i \sim& N(0, \sigma_b^2). \end{aligned}$$

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

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

Notice that if a larger selection trait is associated with a higher probability of selection, then evaluated clusters 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

$$\begin{aligned} P\left(Y_{ij} = 1 | b_i, X_{ij} = 1\right) =& \pi_{ij} \frac{\int_{\gamma_{ij}}^1 \theta_{ij} f(\theta_{ij}) \, d\theta_{ij}}{\int_{\gamma_{ij}}^1 f(\theta_{ij}) \, d\theta_{ij}}, \\ P\left(Y_{ij} = 1 | b_i, X_{ij} = 0\right) =& \pi_{ij} \frac{\int_0^{\gamma_{ij}} \theta_{ij} f(\theta_{ij}) \, d\theta_{ij}}{\int_0^{\gamma_{ij}} f(\theta_{ij}) \, d\theta_{ij}}. \end{aligned}$$

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, under monotonicity, the use of likelihood (6) in combination with (2) will produce an upper bound for the probabilities of success in presence of selection bias. 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}.

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 out 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

$$\begin{aligned} L_c^*(\boldsymbol{\beta}, \lambda, \tau, b_i) &= \int f\left(\boldsymbol{Y}_{1i}, \boldsymbol{\theta}_{1i}, b_i\right)\, d\boldsymbol{\theta}_{1i}, \\ &= \prod_{j=1}^{N_i} \left\{ \frac{1}{\lambda+\tau} (\pi_{ij}\lambda)^{y_{ij}} \left[(1-\pi_{ij})\lambda + \tau\right]^{1-y_{ij}} \right\}, \end{aligned} \tag{10}$$

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

$$L_m(\boldsymbol{\beta}, \lambda, \tau, \sigma^2) = \prod_i^n \int L_c^*(\boldsymbol{\beta}, \lambda, \tau, b_i) \phi(b_i|0, \sigma^2)\, db_i. \tag{11}$$

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 model (1) could be fitted using maximum likelihood. This minimizes the numerical noise and provides a clearer idea regarding the performance of both approaches. Two hundred data sets were generated,

with the following parameters held constant across data sets: (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: (1) the number of ratings per expert N_i and (2) a set of 147 expert random-effects b_i. The random rater specific effects b_i were sampled from $N(0, 10)$, and in the original data we assumed that all experts rated all clusters, i.e., $N_i = N = 30$. The actual clusters evaluated by each rater (N_i) were then defined using the selection process $X_{ij}|b_i \sim$ Bernoulli (ρ_{ij}) with logit$(\rho_{ij}) = \alpha_j + b_i$. 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.

For the selection probabilities two settings were consider. In a first scenario selection bias was not present and the ratings $Y_{ij}|b_i$ were generated from a Bernoulli(π_{ij}) with

$$\pi_{ij} = \frac{\exp(\beta_j + b_i)}{1 + \exp(\beta_j + b_i)} .$$

In the second scenario there was a selection bias in the rating process and the rating probabilities were generated as

$$\text{logit}\left[P(Y_{ij} = 1|X_{ij} = x_{ij}, b_i)\right] = \begin{cases} \beta_j + b_i & \text{if } x_{ij} = 1, \\ \beta_j + b_i - 0.223 & \text{if } x_{ij} = 0. \end{cases} \tag{12}$$

Basically, (12) 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.

Notice that the scenarios used in the simulations are a special case of the general modeling framework introduced in Section 3. In fact, to simplify the computational burden and improve numerical stability, we considered the situation in which the selection and rating procedures shared a common random effect. This is the so-called shared parameter model (SPM), for which corr$(a_i, b_i) = 1$ [8].

5.1 Results in absence of selection bias

In this scenario three analyses were carried out for each data set and the main results are summarized in tables 1–4. In these tables, the column *True* gives always the true value of the corresponding parameter, the column *Combined* refers to results obtained from the combined model introduced in Section 4, the column J(·) displays the results obtained from fitting the

joint model using the selection probability derived from the logit in brackets and, finally, the column *Naive* presents the results obtained from fitting model 1 without accounting for the different selection probabilities.

Notice that model J$(\alpha_j + b_i)$ assumes that the selection probabilities vary across clusters for each rater, but the parameters governing the rating and selection processes are different. In the present simulations this model correctly described the data generating mechanism. 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 a specific expert rated.

The combined model is misspecified in this scenario since the selection model used in generating the data is not equivalent to the one assumed in the combined model.

Tables 1–2 show that, the combined model performs well when compared to the true values. Unfortunately the correctly specified joint model suffered heavily from lack of convergence problems which led the highly biased point estimates and thus did not offer a good comparison ground against the other models, the true values will mainly be used for comparison. A total of 61 parameters had to be estimated from this model which when compared to the number of experts (147), convergence problems would not be unusual. For smaller number of clusters (15), this model produces almost unbiased point estimates but we preferred to use as many clusters as most to be as close as possible to the case study. On the other hand the bias of the point estimates from the misspecified joint model that assumes equal selection probability for the clusters rated by the same expert, i.e J$(\alpha + b_i)$ was reasonably small, this is a special case of the correctly specified model. Only 32 parameters had to be estimated and thus computationally lighter than the correctly specified model. Nonetheless, for the other misspecified joint model, J$(\beta_j + b_i)$, relative biases larger than 1000% appear. Similar problems occur when the selection process is ignored. Indeed, as the results from the naive analysis show, relative biases larger than 3000% can be obtained when the selection process is incorrectly ignored.

In spite of being misspecified, the combined model always led to unbiased estimates of the parameters as observed in Table 2 . The extra set of random effects in the combined model probably absorbs the extra variation that results from the selection process despite the misspecification of the distribution. Recall that no specific assumptions are made on the mean function of this distribution. The parameters can therefore be estimated such that the resulting shape of the distribution is close to that of the extra variation in the data. While this avoids misspecification of such parameters, the parameters are likely to be estimated with a considerable degree of uncertainty since not a lot of information about them is given beforehand. This is evident in the largest standard errors observed for the estimates from the combined model as seen in Table 1 and the wider confidence interval length in Table 3. On the

other hand, in the misspecified joint model, mean function of the distribution for the selection process is misspecified, this forces estimation of wrong parameters for the selection part which in turn affects estimation of parameters for the rating part since the processes are marginally dependent. As some information for the parameters of the distribution of the variation due to the selection process is provided in this scenario, the wrong estimates are estimated with high precision. As a result, the misspecified joint models exhibited in some settings large bias and high precision, while the combined model had smaller bias and lower precision.

It is important to point out that highly precise but incorrect estimates could lead to seriously misleading inferences. In fact, as shown in Tables 2–3, the fixed effects parameters were estimated with high precision when model J$(\beta_j + b_i)$ was used, however, the confidence interval coverage for eighteen of them was below 50% and it was approximately 0% for thirteen of them. Similarly, the naive model also exhibited a poor performance with coverage probabilities sometimes far below the pre-specified 95%. In contrast, the combined model always produced confidence intervals with good coverage.

Finally Table 4 displays the true and estimated probabilities of success for every compound. Here again the combined model led to estimated values that are almost equal to the true probabilities despite the multiplicative factor in $\frac{\lambda}{\lambda+\tau}$ in (9). This is because the estimated $\tau/\lambda \approx 0$ implying $\frac{\lambda}{\lambda+\tau} \approx 1$, and it can be shown that this corresponds to values of $\theta_{ij} \approx 1$. However, the misspecified and naive models produced biased results with relative biases as large as 40% in some scenarios.

5.2 Results in presence of selection bias

TO BE WRITTEN

6 Case Study Analysis

The case study introduced in Section 2 was analyzed by [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. A summary of the analyses can be found in Table 5 where the cluster are ordered according to the results obtained from the naive model. Remarkably, the three approaches lead to strikingly different results. First, notice that the probabilities of success derived from the joint model are substantially smaller than those obtained from the naive and combined methods. Secondly, the ranks given to the cluster by the three approaches also differ in important ways. For instance, the fourth best compound according to the naive approach (296443) received ranks 911 and 80 from the joint and combined models respectively. Moreover, compound 295061 ranked first by the naive and joint models was not among the top ten cluster

according to the combined model.

Sensitivity of the results with respect to the modeling approach represents a clear dilemma when analyzing this problem. Several strategies could be implemented here, for instance, one could compute the average rank (probability of success) over the different approaches and select those cluster with the largest average rank (probability). On the other hand, given the results of the simulations one could argue that, unlike the naive and joint models, the combined model seems to produced unbiased estimates in most circumstances and, therefore, 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 in a situation like this one. 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.

Given the complexity of the models used in the analyses of the case study and the high dimensionality of the data, the marginal likelihood was computed using the Laplace approximation. Unlike in the case study, the data used in the simulations had a relatively a lower dimension and this allowed to approximate the marginal likelihood using adaptive Gaussian quadrature. It has been shown that these choices may have a non-negligible impact on the results [17]. More complex models are often less biased, but they may require a cruder approximation of the likelihood. Simpler models often allow a better approximation of the likelihood, but they may also be more prone to serious bias. The optimal balance between complexity and precision is difficult to determine in real examples where the true is unknown and this difficulty emphasizes the importance of using all information available when interpreting the results in the decision making process.

7 Conclusion

When quantifying expert opinion in the drug discovery process, one often needs to jointly fit hierarchical models describing the selection and rating mechanisms in order to obtain valid estimates. However, in the present work it has been shown that the joint modeling approach may produce biased estimates of the relevant parameters when selection bias is present and/or the selection model is misspecified.

We have introduced an alternative approach based on the so-called combined model that 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. The loss of precision observed with the combined model may be seen as the price to pay for the robustness achieved.

Given the robustness 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.

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. For instance, 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 overpasses the scope of this work exploring this alternative is certainly worth pursuing.

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

2. Oxman, A.D, Lavis, J.N., and Fretheim, A. . Use of evidence in WHO recommendations. *Lancet* 2007; **369**: 1883-1889.

3. Hack, M.D., Rassokhin, D.N., Buyck, C., Seierstad, M., Skalkin, A., ten Holte, P., Jones, T.K., Mirzadegan, T., and Agrafiotis, D.K. 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. and Bijnens,L. A permutational-splitting sample procedure to quantify expert opinion on chemical cluster using high-dimensional data. *submitted* 2013.

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., Aerts, M., Molenberghs, G., Verbeke, G., and Kenward, M.G. Generalized shared-parameter models and missingness at random. *Statistical Modeling*, 2011 **11**, 279–311.

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

9. Rizopoulos, D., Verbeke, G., and 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, E. F., Green, T., and Schluchter, M. D. 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., and Best, N. Adjusting for selection effects in epidemiologic studies; Why sensitivity analysis is the only "solution". *Commentary in Epidemiology* 2011; **22**: 36-39.

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

14. Molenberghs, G., Verbeke, G., Demétrio, C., and 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, D.K., Alex, S., Dai, H., Derkinderen, A., Farnum, M., Gates, P., Izrailev, S., Jaeger, E. P., Konstant, P., Leung, A., Lobanov,V. S., Marichal, P., Martin, D., Rassokhin, D. N., Shemanarev, M., Skalkin, A., Stong, J., Tabruyn, T., Vermeiren, M., Wan, J., Xu, X. Y., and Yao, X. Advanced Biological and Chemical Discovery (ABCD): Centralizing discovery knowledge in an inherently decentralized world. *J. Chem. Inf. Model*, 2007; **47**, 1999–2014.

16. Frederic, P. and Lad, F. Two Moments of the Logitnormal Distribution. *Communications in Statistics-Simulation and Computation* 2008; **37**: 1263–1269

17. Lesaffre, E. and Spiessens, B. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Appl. Statist.*, 2001; **5**0, 325-335.

Table 1: Point estimates and (standard errors) for different models fitted to data generated under conditional independence assumption. True: true cluster-effect value Comb:combined model cluster-effect estimate and (standard error), $J(\alpha_j + b_i)$:joint mode where $\text{logit}[P(x_{ij} = 1|b_i)] = \alpha_j + b_i$, $J(\beta_j + b_i)$: joint model where $\text{logit}[P(x_{ij} = 1|b_i)] = \beta_j + b_i$, $J(\alpha + b_i)$: joint model where $\text{logit}[P(x_{ij} = 1|b_i)] = \alpha + b_i$, Naive: Naive model cluster-effect estimate. $\text{logit}[P(x_{ij} = 1|b_i)] = \alpha_j + b_i$ was used in selecting the clusters.

β_j	True	Comb		$J(\alpha_j + b_i)$		$J(\beta_j + b_i)$		$J(\alpha + b_i)$		Naive	
β_1	3.60	3.60	(0.93)	6.83	(0.47)	0.20	(0.34)	3.72	(0.76)	5.06	(0.85)
β_2	-1.98	-1.97	(0.82)	-0.15	(0.22)	-0.29	(0.27)	-1.74	(0.36)	-1.09	(0.36)
β_3	4.33	4.33	(0.93)	3.46	(0.59)	2.10	(0.34)	4.32	(0.73)	5.43	(0.84)
β_4	0.58	0.58	(0.77)	2.18	(0.43)	-0.47	(0.32)	0.79	(0.52)	1.77	(0.57)
β_5	0.11	0.11	(0.74)	1.63	(0.36)	-0.49	(0.32)	0.37	(0.47)	1.18	(0.50)
β_6	-0.53	-0.52	(0.75)	1.34	(0.34)	-0.74	(0.31)	-0.29	(0.45)	0.54	(0.47)
β_7	1.70	1.70	(1.03)	2.90	(0.72)	-0.99	(0.34)	1.88	(0.74)	2.91	(0.81)
β_8	-0.10	-0.10	(0.67)	1.34	(0.30)	-0.11	(0.31)	0.03	(0.41)	0.86	(0.43)
β_9	1.51	1.51	(0.89)	2.84	(0.59)	-0.43	(0.33)	1.80	(0.67)	2.79	(0.73)
β_{10}	1.29	1.29	(0.74)	2.66	(0.51)	0.06	(0.32)	1.42	(0.53)	2.33	(0.58)
β_{11}	0.88	0.88	(1.02)	2.74	(0.73)	-1.47	(0.34)	1.11	(0.75)	2.14	(0.79)
β_{12}	-3.52	-3.52	(1.05)	-0.84	(0.26)	-1.19	(0.27)	-3.19	(0.41)	-2.51	(0.41)
β_{13}	0.60	0.60	(0.68)	1.85	(0.37)	-0.03	(0.31)	0.69	(0.46)	1.60	(0.48)
β_{14}	1.89	1.89	(1.06)	3.06	(0.73)	-1.01	(0.34)	2.06	(0.78)	3.18	(0.83)
β_{15}	0.68	0.69	(0.67)	1.55	(0.32)	0.19	(0.31)	0.77	(0.44)	1.68	(0.47)
β_{16}	0.05	0.05	(0.67)	1.18	(0.28)	-0.04	(0.31)	0.14	(0.41)	1.07	(0.43)
β_{17}	0.11	0.12	(0.58)	0.78	(0.21)	1.01	(0.29)	0.22	(0.36)	0.94	(0.36)
β_{18}	-4.01	-4.01	(1.08)	-1.28	(0.26)	-1.05	(0.25)	-3.74	(0.42)	-3.06	(0.42)
β_{19}	0.27	0.27	(0.67)	1.35	(0.29)	0.19	(0.31)	0.40	(0.41)	1.23	(0.43)
β_{20}	-1.79	-1.79	(0.77)	-0.11	(0.21)	-0.03	(0.27)	-1.57	(0.36)	-0.92	(0.35)
β_{21}	1.03	1.03	(0.66)	1.60	(0.31)	0.50	(0.31)	1.10	(0.45)	1.99	(0.47)
β_{22}	0.03	0.03	(0.70)	1.38	(0.32)	-0.34	(0.31)	0.19	(0.44)	1.08	(0.47)
β_{23}	-0.05	-0.05	(0.72)	1.47	(0.34)	-0.34	(0.31)	0.15	(0.44)	1.04	(0.46)
β_{24}	-0.95	-0.95	(0.71)	0.43	(0.22)	0.00	(0.28)	-0.81	(0.37)	-0.07	(0.37)
β_{25}	0.10	0.10	(0.75)	1.66	(0.37)	-0.57	(0.32)	0.38	(0.48)	1.22	(0.51)
β_{26}	0.87	0.88	(0.68)	1.78	(0.34)	0.27	(0.32)	0.98	(0.46)	1.86	(0.48)
β_{27}	2.13	2.13	(1.24)	6.42	(0.65)	-1.40	(0.35)	2.55	(0.81)	3.60	(0.89)
β_{28}	-3.03	-3.03	(0.94)	-0.79	(0.23)	-0.69	(0.26)	-2.80	(0.38)	-2.20	(0.38)
β_{29}	-0.34	-0.33	(0.68)	0.92	(0.25)	0.11	(0.30)	-0.16	(0.39)	0.63	(0.40)
β_{30}	2.06	2.06	(0.85)	2.84	(0.57)	0.31	(0.33)	2.02	(0.62)	3.13	(0.69)
$\widehat{\sigma}^2$	10.00	10.16	(6.97)	10.42	(1.53)	7.81	(1.12)	7.81	(1.12)	5.80	(1.19)

Table 2: Relative bias for different models fitted to data generated under conditional independence assumption. Comb: combined model,$J(\alpha_j + b_i)$:joint mode where $\text{logit}[P(x_{ij} = 1|b_i)] = \alpha_j + b_i$, $J(\beta_j + b_i)$: joint model where $\text{logit}[P(x_{ij} = 1|b_i)] = \beta_j + b_i$, $J(\alpha+b_i)$: joint model where $\text{logit}[P(x_{ij} = 1|b_i)] = \alpha+b_i$, Naive: Naive model cluster-effect estimate.$\text{logit}[P(x_{ij} = 1|b_i)] = \alpha_j + b_i$ was used in selecting the clusters.

β_j	Comb	$J(\alpha_j + b_i)$	$J(\beta_j + b_i)$	$J(\alpha + b_i)$	Naive
β_1	0.00	0.90	0.94	0.03	0.40
β_2	0.00	0.93	0.85	0.12	0.45
β_3	0.00	0.20	0.52	0.00	0.25
β_4	0.00	2.74	1.80	0.35	2.04
β_5	0.02	14.28	5.63	2.48	10.02
β_6	0.01	3.55	0.40	0.45	2.03
β_7	0.00	0.71	1.58	0.11	0.71
β_8	0.03	14.31	0.06	1.28	9.56
β_9	0.00	0.89	1.28	0.20	0.85
β_{10}	0.00	1.06	0.95	0.10	0.80
β_{11}	0.00	2.12	2.68	0.27	1.44
β_{12}	0.00	0.76	0.66	0.10	0.29
β_{13}	0.00	2.07	1.05	0.14	1.66
β_{14}	0.00	0.62	1.54	0.09	0.68
β_{15}	0.00	1.26	0.72	0.12	1.46
β_{16}	0.06	22.19	1.75	1.68	19.96
β_{17}	0.03	5.85	7.89	0.97	7.20
β_{18}	0.00	0.68	0.74	0.07	0.24
β_{19}	0.01	4.06	0.29	0.51	3.64
β_{20}	0.00	0.94	0.98	0.12	0.48
β_{21}	0.00	0.55	0.51	0.06	0.93
β_{22}	0.09	51.51	14.02	6.24	39.83
β_{23}	0.06	31.51	6.00	4.11	22.56
β_{24}	0.00	1.45	1.00	0.15	0.92
β_{25}	0.02	15.20	6.61	2.69	10.93
β_{26}	0.00	1.04	0.69	0.12	1.13
β_{27}	0.00	2.02	1.66	0.20	0.69
β_{28}	0.00	0.74	0.77	0.08	0.28
β_{29}	0.01	3.74	1.32	0.54	2.88
β_{30}	0.00	0.38	0.85	0.02	0.52
σ	0.02	0.04	0.22	0.22	0.42

Table 3: Confidence interval coverage and (length of confidence interval) for different models fitted to data generated under conditional independence assumption. Comb: combined model, $J(\alpha_j + b_i)$:joint mode where $\text{logit}[P(x_{ij} = 1|b_i)] = \alpha_j + b_i$, $J(\beta_j + b_i)$: joint model where $\text{logit}[P(x_{ij} = 1|b_i)] = \beta_j + b_i$, $J(\alpha + b_i)$: joint model where $\text{logit}[P(x_{ij} = 1|b_i)] = \alpha + b_i$, Naive: Naive model cluster-effect estimate.$\text{logit}[P(x_{ij} = 1|b_i)] = \alpha_j + b_i$ was used in selecting the clusters.

β_j	Comb		$J(\alpha_j + b_i)$		$J(\beta_j + b_i)$		$J(\alpha + b_i)$		Naive	
β_1	0.97	(15.18)	0.50	(2.10)	0.00	(1.32)	0.60	(2.85)	0.82	(3.67)
β_2	0.92	(10.32)	0.06	(0.90)	0.00	(1.06)	0.91	(1.43)	0.31	(1.44)
β_3	0.96	(9.01)	0.72	(2.64)	0.00	(1.34)	0.87	(2.99)	0.93	(3.59)
β_4	0.92	(8.34)	0.06	(1.81)	0.06	(1.27)	0.98	(2.12)	0.48	(2.31)
β_5	0.91	(8.01)	0.06	(1.48)	0.54	(1.25)	0.94	(1.92)	0.42	(2.02)
β_6	0.93	(8.95)	0.06	(1.36)	0.90	(1.23)	0.90	(1.78)	0.34	(1.87)
β_7	0.97	(9.59)	0.56	(2.77)	0.00	(1.32)	0.80	(2.90)	0.80	(3.30)
β_8	0.91	(7.41)	0.06	(1.21)	0.96	(1.20)	0.95	(1.63)	0.37	(1.71)
β_9	0.96	(7.66)	0.39	(2.14)	0.00	(1.30)	0.93	(4.57)	0.66	(3.03)
β_{10}	0.95	(7.50)	0.33	(2.00)	0.01	(1.26)	0.97	(2.18)	0.57	(2.38)
β_{11}	0.96	(9.18)	0.06	(2.92)	0.00	(1.33)	0.90	(3.57)	0.78	(3.28)
β_{12}	0.93	(14.03)	0.06	(1.05)	0.00	(1.04)	0.85	(1.61)	0.35	(1.62)
β_{13}	0.93	(6.68)	0.06	(1.53)	0.49	(1.23)	0.97	(1.81)	0.44	(1.95)
β_{14}	0.97	(9.42)	0.89	(3.16)	0.00	(1.34)	0.81	(3.05)	0.79	(3.44)
β_{15}	0.94	(6.74)	0.11	(1.35)	0.63	(1.23)	0.95	(1.76)	0.42	(1.88)
β_{16}	0.95	(7.63)	0.00	(1.13)	0.97	(1.20)	0.94	(1.63)	0.40	(1.71)
β_{17}	0.92	(6.36)	0.06	(0.90)	0.11	(1.13)	0.97	(1.40)	0.39	(1.43)
β_{18}	0.95	(14.85)	0.11	(1.10)	0.00	(0.98)	0.88	(1.65)	0.41	(1.66)
β_{19}	0.92	(5.87)	0.06	(1.23)	0.97	(1.21)	0.95	(1.65)	0.36	(1.73)
β_{20}	0.92	(9.73)	0.00	(0.86)	0.00	(1.05)	0.93	(1.40)	0.35	(1.40)
β_{21}	0.92	(6.12)	0.50	(1.28)	0.63	(1.23)	0.95	(1.77)	0.49	(1.90)
β_{22}	0.92	(7.50)	0.06	(1.28)	0.80	(1.23)	0.92	(1.76)	0.43	(1.85)
β_{23}	0.93	(7.12)	0.06	(1.33)	0.88	(1.22)	0.97	(1.74)	0.35	(1.85)
β_{24}	0.92	(8.70)	0.06	(0.93)	0.09	(1.12)	0.92	(1.44)	0.36	(1.47)
β_{25}	0.92	(8.09)	0.06	(1.47)	0.46	(1.26)	0.94	(1.90)	0.37	(2.03)
β_{26}	0.94	(6.99)	0.11	(1.48)	0.50	(1.24)	0.96	(1.82)	0.49	(1.96)
β_{27}	0.96	(10.88)	0.50	(2.18)	0.00	(1.38)	0.61	(4.86)	0.62	(3.33)
β_{28}	0.96	(12.50)	0.06	(0.96)	0.00	(1.01)	0.91	(1.51)	0.40	(1.52)
β_{29}	0.92	(7.29)	0.11	(1.05)	0.66	(1.17)	0.95	(1.52)	0.29	(1.57)
β_{30}	0.97	(7.41)	0.89	(2.18)	0.00	(1.28)	0.98	(2.58)	0.78	(2.89)
σ	0.89	(94.32)	0.94	(6.27)	0.51	(4.47)	0.49	(4.49)	0.19	(4.74)

Table 4: Probability estimates and (relative bias) ranked from the highest to the lowest for different models fitted to data generated under conditional independence assumption. True: true cluster-effect value Comb: combined model cluster-effect estimate and (standard error), $J(\alpha_j + b_i)$:joint mode where $\text{logit}[P(x_{ij} = 1|b_i)] = \alpha_j + b_i$, $J(\beta_j + b_i)$: joint model where $\text{logit}[P(x_{ij} = 1|b_i)] = \beta_j + b_i$, $J(\alpha + b_i)$: joint model where $\text{logit}[P(x_{ij} = 1|b_i)] = \alpha + b_i$, Naive: Naive model cluster-effect estimate. $\text{logit}[P(x_{ij} = 1|b_i)] = \alpha_j + b_i$ was used in selecting the clusters.

β_j	True	Comb		$J(\alpha_j + b_i)$		$J(\alpha + b_i)$		$J(\beta_j + b_i)$		Naive	
β_3	0.88	0.88	(0.00)	0.83	(0.06)	0.74	(0.17)	0.90	(0.02)	0.96	(0.09)
β_1	0.84	0.84	(0.00)	0.94	(0.12)	0.52	(0.38)	0.87	(0.04)	0.95	(0.13)
β_{27}	0.72	0.72	(0.00)	0.92	(0.28)	0.34	(0.53)	0.78	(0.08)	0.88	(0.23)
β_{30}	0.72	0.71	(0.00)	0.78	(0.09)	0.54	(0.25)	0.73	(0.02)	0.85	(0.19)
β_{14}	0.70	0.70	(0.00)	0.81	(0.16)	0.38	(0.46)	0.73	(0.04)	0.85	(0.22)
β_7	0.68	0.68	(0.00)	0.79	(0.16)	0.38	(0.44)	0.72	(0.05)	0.84	(0.23)
β_9	0.66	0.66	(0.00)	0.78	(0.18)	0.45	(0.32)	0.70	(0.06)	0.82	(0.25)
β_{10}	0.64	0.64	(0.00)	0.75	(0.18)	0.51	(0.21)	0.67	(0.05)	0.79	(0.23)
β_{21}	0.61	0.61	(0.00)	0.67	(0.09)	0.56	(0.08)	0.63	(0.03)	0.75	(0.22)
β_{11}	0.60	0.60	(0.00)	0.78	(0.30)	0.33	(0.45)	0.63	(0.05)	0.76	(0.28)
β_{26}	0.60	0.59	(0.00)	0.68	(0.15)	0.53	(0.10)	0.62	(0.04)	0.73	(0.22)
β_{15}	0.57	0.57	(0.00)	0.67	(0.16)	0.52	(0.09)	0.59	(0.03)	0.71	(0.23)
β_{13}	0.57	0.57	(0.00)	0.69	(0.21)	0.50	(0.12)	0.58	(0.03)	0.71	(0.25)
β_4	0.56	0.56	(0.00)	0.72	(0.28)	0.44	(0.21)	0.59	(0.06)	0.72	(0.28)
β_{19}	0.53	0.53	(0.00)	0.65	(0.22)	0.52	(0.01)	0.55	(0.04)	0.66	(0.25)
β_{17}	0.51	0.51	(0.00)	0.59	(0.14)	0.62	(0.21)	0.53	(0.03)	0.62	(0.22)
β_5	0.51	0.51	(0.00)	0.67	(0.32)	0.44	(0.14)	0.54	(0.06)	0.66	(0.28)
β_{25}	0.51	0.51	(0.00)	0.68	(0.32)	0.43	(0.16)	0.55	(0.07)	0.66	(0.29)
β_{16}	0.51	0.51	(0.00)	0.62	(0.23)	0.50	(0.02)	0.52	(0.02)	0.64	(0.26)
β_{22}	0.50	0.50	(0.00)	0.64	(0.27)	0.46	(0.09)	0.52	(0.04)	0.64	(0.27)
β_{23}	0.49	0.49	(0.00)	0.66	(0.33)	0.46	(0.07)	0.52	(0.05)	0.63	(0.28)
β_8	0.49	0.49	(0.00)	0.64	(0.31)	0.49	(0.01)	0.50	(0.03)	0.62	(0.26)
β_{29}	0.46	0.46	(0.00)	0.60	(0.30)	0.51	(0.11)	0.48	(0.04)	0.59	(0.26)
β_6	0.44	0.44	(0.00)	0.64	(0.45)	0.41	(0.07)	0.47	(0.05)	0.57	(0.30)
β_{24}	0.40	0.40	(0.00)	0.55	(0.38)	0.50	(0.26)	0.40	(0.02)	0.49	(0.24)
β_{20}	0.31	0.31	(0.00)	0.49	(0.57)	0.50	(0.60)	0.32	(0.03)	0.38	(0.22)
β_2	0.29	0.29	(0.00)	0.48	(0.65)	0.46	(0.59)	0.30	(0.02)	0.36	(0.23)
β_{28}	0.20	0.20	(0.01)	0.41	(1.05)	0.42	(1.07)	0.20	(0.02)	0.23	(0.17)
β_{12}	0.17	0.17	(0.01)	0.41	(1.47)	0.36	(1.17)	0.17	(0.01)	0.20	(0.20)
β_{18}	0.13	0.14	(0.01)	0.36	(1.66)	0.38	(1.80)	0.13	(0.03)	0.15	(0.14)

Table 5: *Estimated parameters ($\hat{\beta}$), probabilities of success ($\widehat{P}$) and ranks for the top 20 clusters (according to the naive 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	$\hat{\beta}$	$\widehat{\text{P}}$	Rank	$\hat{\beta}$	$\widehat{\text{P}}$	Rank	$\hat{\beta}$	$\widehat{\text{P}}$	Rank
265222	2.52	0.94	1	2.67	0.72	3	0.97	0.62	25
295061	3.83	0.92	2	2.61	0.71	4	1.69	0.71	1
359957	0.49	0.87	3	-0.25	0.48	330	0.71	0.59	88
69850	1.07	0.82	4	0.11	0.50	182	0.89	0.61	38
84163	5.24	0.77	5	1.83	0.65	9	1.34	0.67	6
296443	2.59	0.76	6	1.62	0.64	10	0.55	0.57	162
7451	1.28	0.74	7	0.66	0.56	55	0.61	0.57	147
277619	1.65	0.73	8	0.44	0.54	89	0.60	0.58	138
315928	2.04	0.72	9	1.47	0.62	14	1.26	0.66	9
296535	2.77	0.71	10	2.37	0.70	5	1.58	0.70	2
313914	2.18	0.70	11	2.06	0.68	7	1.47	0.68	4
277774	2.20	0.69	12	1.30	0.61	20	0.98	0.62	24
178994	1.85	0.68	13	1.57	0.64	11	1.14	0.65	13
296560	1.89	0.66	14	1.86	0.66	8	1.09	0.64	15
464822	1.21	0.66	15	0.56	0.55	72	0.90	0.61	40
265441	1.87	0.65	16	1.44	0.62	15	0.90	0.61	34
292805	1.47	0.65	17	1.20	0.61	19	1.06	0.64	20
432169	1.45	0.64	18	6.26	0.91	1	1.01	0.63	21
292579	1.85	0.64	19	1.50	0.63	13	1.23	0.65	11
278927	1.30	0.63	20	0.51	0.54	76	0.97	0.62	26
σ^2	20.02			18.61			6.64		