2013•2014
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Modeling of selection process during evaluation of clusters chemical of chemical compounds in the drug discovery process

Promotor :
dr. Elasma MILANZI

David Amwonya
Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2013•2014
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Modeling of selection process during evaluation of clusters chemical of chemical compounds in the drug discovery process

Promotor :
dr. Elasma MILANZI

David Amwonya
Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

universiteit hasselt | Maastricht University

Modeling the selection process during evaluation of clusters of chemical compounds in the drug discovery process

By

David Amwonya

Supervisor:

Dr. Elasma Immaculate Milanzi

February 05, 2014

A thesis submitted to the Center for Statistics, Universiteit Hasselt, Belgium in Partial Fulfilment of the the Requirement for the Degree of Master of Science in Biostatistics.

Certification

I declare that this thesis was written, by me under the guidance and counsel of my supervisor.

. . . .

David Amwonya **Date**

I certify that this is the true report written by **David Amwonya** under my supervision and thus I permit its presentation for assessment.

. . . .

Dr. Elasma Immaculate Milanzi **Date**

Acknowledgement

Successful completion of this thesis has been through the support of my supervisor Dr. Elasma Immaculate Milanzi who continually guided me through this work, by supplying thoughtful suggestions to improve on my report. I also send my gratitude to VLIR-UOS scholarship for awarding me this important opportunity for developing my knowledge and expertise in the field of statistics which has really improved my views in terms of my academic perspective and statistics as a whole. Thanks to VLIR-UOS coordinator in Hasselt University Mrs. Martine Machiels and Mr. Marc Thoelen for all the support and timely communication during the entire study period. I thank all the professors of the master of statistics programme for their guidance and teaching expertise.

I am also grateful to my Landlord Luc Vern Herck, my friend Blessing, the Ugandan friends: Abdallah, Ashraf, Lawrence, Julius, Mary, Patrick, Edward, Geraldine, Swabra, Gorge, Suzan, Tony, Henry, Doreen, all my classmates and friends not forgetting my Campus Pastor Karin and CCg members for making my stay in Belgium lively and enjoyable.

To my parents, brothers, sisters, my great friends Dr. Rebecca Nsubuga and Peluce thank you for your words of encouragement.

Most of all I thank God for giving me wisdom, strength and life to successfully complete this program.

Abstract

Drug development is a lengthy and expensive process which undergoes several stages. To reduce on the time and money, pharmaceutical companies tend to maintain a library of chemical compounds that are screened for some drug-like activities. To increase on the number of hits in screening process large volumes of compounds are required which can be got from vendors. However, this come with challenges on evaluation of compounds worth purchasing hence experts are employed to help in the correct evaluation process. Since the process is lengthy and tiresome, we therefore need to come-up with proper evaluation procedure for example letting expert evaluate cluster of compounds based on their appropriate time schedule. However, such a procedure can introduce selection bias. Therefore the main gist of the project was to explore an alternative method to model the selection bias through use of extra set of random effects. To achieve this, we employed three modelling techniques; the 'naive model' which ignores the selection process, the 'joint model' which takes into account the selection process and then the 'combined model' which uses extra set of random effects as an alternative approach to the joint model. The results showed that ignoring the selection process had a bigger impact on the estimates, misspecifying the selection (joint) model introduces severe bias in the estimates of the parameters. The combined model seemed to be a robust method as its results were closely related to the true value and the correctly specified joint model. Therefore one would use the combined model as a tool for sensitivity analysis and alternative method in making inferences in situations where the joint model is not working properly.

Key words: Combined model, Joint model, Naive model, Selection bias .

Contents

List of Tables

1 Introduction

Drug development is a lengthy and expensive process which undergoes several stages like discovery, optimization and development stage (Paul *et al.,* 2010). It involves identification of protein targets and ends after clinical trials, after which similar compounds in terms of properties are grouped into clusters that are qualitatively assessed. To minimise the time and money spent in this process, pharmaceutical companies tend to maintain a library of chemical compounds that are screened for some drug-like activities. To increase the chances of hits during the screening process, it is advisable to have atleast large proportion of interetsing compounds (Lajiness & Watson, 2000). This therefore requires acquisition of more chemical compounds from the vendors. However, this comes with challenges such as evaluating what compounds are worth purchasing. As a result expert knowledge is sought to carefully evaluate the amount of evidence supporting potential compounds before investing more resources in them (Alonso & Molenberghs, 2008).

Since the process of chemical compound acquisition involves large number of clusters of compounds, there is need to come-up with an appropriate evaluation procedure. Most practical proposal is to allow experts decide on what to evaluate depending on the experts' schedule, compelling the experts to evaluate all clusters though this might not be feasible in reality. Such a procedures were used in the study that motivated this work and was conducted by Janssen pharmaceutical, but due to some legal implications we will not discuss the case study, we will instead work with simulated data. Another method could be assigning a single subset of clusters randomly to experts to evaluate and then compelling them to finish their quota (Alonso *et al,* 2013). Alternatively, automated programmed machines (computers) that mimic the expert opinions can be used to carry out the evaluation process though similar challenges as above are encountered (Kearsley *et al.,*

1996 and Sheridan *et al.,* 1996). Therefore appropriate statistical methods are required to quantify the success probability for each cluster, where 'success' refers to recommending the inclusion of a cluster into the sponsors database(library) for future scrutiny (Milanzi, 2013 and Milanzi *et al.,* 2013).

Alonso *et al.,* (2013) showed that some evaluation 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. The above authors demonstrated that even in absence of selection bias, one often needs to jointly model the rating and selection processes in order to avoid bias. 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.

To account for the selection process, we have used the 'joint model' where two sets of Generalized Linear Mixed Models (GLMM) are employed. It assumes conditional independence of the selection and the rating process based on some random effects. However, the two models themselves are marginally dependent. This conditional independence assumption in this model is closely related to the Shared Parameter (SP) and Generalized Shared Parameter (GSP) modelling frameworks used to describe a Missing Not At Random (MNAR) mechanism in missing data analysis (Creemers *et al.,* 2011 ; Follmann & Wu, 1995). The conditional independence assumption simplifies the joint distribution of the rating and selection processes hence facilitating the joint fit of both models. However, studies have showed that if misspecified, this model might lead to bias and hence inference based on the fixed effects, probabilities e.t.c will greatly be affected. Therefore some authors like Genelletti(2011) advise that a sensitivity analysis to assess the stability of the

results is always appropriate.

Due to the drawbacks of the joint model, we have used another alternative method the 'combined model' to account for selection process. This model was introduced by (Booth *et al.*, 2003 and Molenberghs *et al.,* 2010). It is a member of the exponential family where an extra set of random effects is used to account for overdispersion in correlated outcomes. The advantage with the combined model is that it is only based on the distribution of the random effects without more assumptions hence making it robust against misspecification. For that reason, it can be used as a tool for sensitivity analysis in the presence of selection bias.

Lastly a method which does not take into account the selection process was used. This model will be called the 'Naive model'(mixed logistic regression) and it is simple to fit using GLMM.

The Statistical analysis was performed using the SAS, version 9.2 and R version 2.13.2. All statistical tests were done at 5% level of significance unless stated otherwise.
The report is organised as follows: section 1 contains the introduction and the objective of the study. Section 2 we capture the methodology of estimating the probability of success, the various modelling techniques and simulation study. Then section 3, 4 & 5, we present the results, discussion and conclusion respectively.

1.1 Study objective

To explore whether the combined model can act as an alternative method to account for the selection process by use of an extra set of random effects. Then also to specifically check whether:

- Assuming the selection process has no impact on the estimated probability of recommendation and can therefore be ignored.
- Jointly modelling the evaluation and selection processes by formulating a parametric models for each of the processes.

2 Methodology

2.1 Joint model approach

To be able to capture the information arising from selection and rating process, Milanzi (2013), Milanzi *et al.,* (2013) and Alonso *et al.,* (2013) proposed to summarize the large number of qualitative assessments given by the experts into a single probability of success for every cluster by using a joint model.

Let the vector rating associated with the expert i be $\boldsymbol{Y_i} = (Y_{ij})_{\ j} \in \Lambda_i$ where $\boldsymbol{\Lambda_i}$ is a subset of all clusters evaluated by the i^{th} expert, $i = 1, ..., n$ and $\boldsymbol{Y_i}$ takes on values 1 if liked and 0 otherwise.

Under the conditional independence assumption the above authors showed that the joint distribution of (Y_i, X_i, a_i, b_i) takes the form

$$P(\boldsymbol{Y_i} = y_i, \boldsymbol{X_i} = x_i, a_i, b_i|\boldsymbol{\beta, \alpha, \Sigma}) = P(Y_i = y_i|X_i = x_i, b_i, \beta)P(X_i = x_i|a_i, \beta, \alpha)\phi(a_i, b_i|\boldsymbol{0, \Sigma}) \tag{1}$$

where $\boldsymbol{X_i} = (X_{i1}, ..., X_{iN})$ is the vector containing the selection indicators for expert i and $X_{ij} = 1$ if expert i evaluates cluster j, 0 otherwise. $\boldsymbol{\beta} = (\beta_1, ..., \beta_N)^T$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)^T$ are vectors of parameters due to rating and selection respectively. The a_i and b_i are random selection and rater specific effects respectively. Then N_i is the number of ratings per i^{th} expert and $\boldsymbol{\Sigma}$ is the Hessian matrix obtained from fitting the model.

So we have the distribution of the rating process given by:-

$$P(\boldsymbol{Y_{1i}} = \boldsymbol{y_{1i}}|b_i, \boldsymbol{\beta}) = \prod_{j}^{N_i} P(Y_{1ij} = y_{1ij}|b_i, \beta_j)$$

and the selection process given by:-

$$P(\boldsymbol{X_i} = \boldsymbol{x_i}|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{j}^{N} P(X_{ij} = x_{ij}|a_i, \beta_j, \alpha_j)$$

In the most general scenario, the potential of cluster j can be quantified as

$$P(Y_j = 1) = \int \int P(Y_{ij=1|a_i,b_i})\phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma})da_i db_i \tag{2}$$

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

$$P(Y_{ij} = 1|a_i b_i) = E_x[P(Y_{ij} = 1|X_{ij} = x_{ij}, b_i)]$$

$$= P(Y_{ij} = 1|X_{ij} = 1, b_i)P(X_{ij} = 1|a_i) + P(Y_{ij} = 1|X_{ij} = 0, b_i)P(Xij = 0|a_i)$$

We now have the 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. We also have the information about which clusters every expert evaluated and will use this information to estimate $P(X_{ij} = 1|a_i)$. However, in reality we don't have enough information on $P(Y_{ij} = 1|X_{ij} = 0, b_i)$ therefore we need to make assumptions about it and if these assumptions are wrong, estimates of the rating process will be affected.

The joint likelihood emerging from (1) is given by:-

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \prod_{i}^{n} P(Y_{1i} = y_{1i}, X_i = x_i) | \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \tag{3}$$

Then using the maximum likelihood on (3) estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\Sigma}}$ can be obtained. Then one can estimate the probabilities of success by substituting the estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\sigma}}$ into (2).

Alonso *et al.,* (2013) 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 $P(X_{ij} = x_{ij}|a_i, \alpha_i)$ and $P(Y_{ij} = y_{ij}|X_{ij} = x_{ij}, b_i, \beta_j)$ 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 $(\boldsymbol{a_i}, \boldsymbol{b_i})^T$ are independent and follow a bivariate normal distribution with mean zero and covariance matrix . These authors stated that there is selection bias in the rating process if $P(Y_{ij} = y_{ij}|X_{ij} = x_{ij}, a_i, b_i) \neq P(Y_{ij} = y_{ij}|X_{ij} = i, b_i)$ and showed that absence of selection bias is equivalent to the validity of the following conditional independence assumption:

$$P(Y_{ij} = y_{ij}, X_{ij} = x_{ij}|a_i, b_i) = P(X_{ij} = x_{ij}|a_i) \tag{4}$$

Essentially, (4) states that for every expert, the rating and selection procedures are independent and governed by different, although possibly correlated, random effects.

2.2 Combined model approach

The proposed solution, the combined model was introduced by Booth *et al.*, (2003) and Molenberghs *et al.,* (2010) for members of the exponential family, where an extra set of random effects is used to account for overdispersion in correlated outcomes. In this report,

we shall take into account the selection process by adding a new set of random effects to the rating model. Although the combined model has been shown to improve model fit in overdispersed correlated data, its usefulness to account for selection bias is yet to be investigated.

The combined model follows a different path for estimating the parameters of interest, namely, the fixed specific cluster effects β_j^s and and the inter-expert variability σ^2 . In the joint model, the selection process $P(X_{ij} = x_{ij}|a_i, \beta_i, \alpha_i)$ appearing in the integrand (1) is directly modelled using for instance, a GLMM. We now propose to account for the extra variability emanating from the selection process by introducing a new set of random effects θ_{ij} which is a latent selection trait for every expert . Essentially, we propose to work with the conditional distribution given by

$$f(\boldsymbol{Y_i} = \boldsymbol{y_i}, \boldsymbol{\theta_i}|b_i) = P(\boldsymbol{Y_i} = \boldsymbol{y_i}|b_i, \boldsymbol{\theta_i})f(\boldsymbol{\theta_i}|b_i) = \prod_j^N P(Y_{ij} = y_{ij}|b_i, \theta_{ij})f(\theta_{ij}|b_i) \qquad (5)$$

where $\boldsymbol{\theta_i} = (\theta_{i1}, ..., \theta_{iN})^T$ and $N = (N_1 + N_2 + ... + N_s)$ is the number of clusters rated with S the total number of subsets. The previous expression assumes that, conditional on the random effects, the ratings are independent and so are components of $\boldsymbol{\theta_i}$. Since the two sets of random effects are meant to explain different sources of variability, $\boldsymbol{\theta_i}$ and b_i are also assumed to be independent, hence, $f(\boldsymbol{\theta_i}|b_i) = f(\boldsymbol{\theta_i})$. Finally, $\theta_i \sim Beta(\lambda, \tau)$, $b_i \sim N(0, \sigma^2)$ and $Y_{ij}|\theta_i, b_i \sim Bernoulli(\theta_{ij}, \pi_{ij})$ with $\pi_{ij} = \frac{exp(\beta_j + b_i)}{1+exp(\beta_j + b_i)}$.
Then the probability of success for the cluster of compounds C_j can be calculated by:

$$P(Y_j = 1) = \int\int P(Y_{ij} = 1, \theta_{ij}, b_i)d\theta_{ij}db_i = \frac{\lambda}{\lambda + \tau}E_b(\pi_{ij}) \qquad (6)$$

The validity of the results obtained in the combined model, are based on untestable as-

sumptions like the multiplicative effect of θ_{ij} on π_{ij} and the use of a convenient conjugate distribution of θ_{ij}. Essentially, the strength of the combined model approach lies in using two sets of random effects, one of which is conjugate to the distribution of the ratings. The conjugate random effects account for the selection process, whereas the normal random effect accounts for the correlation within the set of ratings of a given rater. Often the selection process is not of particular scientific interest and does not need to be exhaustively modelled.

2.3 Naive estimation of probability

Naive model was also used to summarise the information arising from the expert evaluation. To estimate the parameter of interest, is equivalent to maximising the likelihood function of the joint model but ignoring the selection process.

2.4 Simulation Study

When working with hierarchical models one often has to deal with likelihood functions that do not have a closed form. For instance, combining normal random effects and binary outcomes with logit links leads to an unclosed form for the marginal likelihood and, therefore, one needs to resort to numerical algorithms to compute the Maximum Likelihood Estimators (MLE). Consequently, studying the properties of the MLE theoretically is extremely difficult in many settings and simulation studies become an indispensable tool to compare alternative approaches in these scenarios (Alonso *et al.,* 2013).

For this project however, to generate the datasets to be used in the analysis, a simulation study was set-up. To that effect, 147 experts and 15 clusters were considered for the

simulations. The fixed effects β_j & α_j,were sampled once from a N(0, 2) and N(0, 1) respectively and then held fixed throughout all simulations, whereas the random rater and selection specific effects b_i and a_i respectively were sampled from a bivariate normal distribution with $mean = [0, 0]$ and variance-covariance matrix

$$M = \begin{bmatrix} 10 & 4.95 \\ 4.95 & 5 \end{bmatrix}$$

respectively. Then the clusters evaluated by each rater were defined using the selection process $X_{ij}|a_i \sim Bernoulli(\rho_{ij})$ with $logit(\rho_{ij}) = \alpha_j + a_i$ and the corresponding rating given by $Y_{ij}|b_i$ were generated from a $Bernoulli(\pi_{ij})$ with $\pi_{ij} = \frac{exp(\beta_j + b_i)}{1+exp(\beta_j + b_i)}$. In this setting a total of 200 data sets were generated.

Then analysis of the three modelling techniques were carried out for each dataset.

3 Results

Table 1: *Estimates (standard errors) for the parameters governing the rating process obtained from the different models fitted to the data. The data were generated by a joint model with selection probability, $logit(\rho_{ij}) = \alpha_j + a_i$. The column True gives the true values, Combined gives the results obtained from the combined model, 'J' refers to the results obtained from a joint model with the logit of the selection probabilities given in brackets and Naive indicates results from the model that disregards the selection process.*

$\boldsymbol{\beta_j}$	**True**	**Combined**	$\boldsymbol{J(\alpha_j + a_i)}$	$\boldsymbol{J(\beta_j + a_i)}$	$\boldsymbol{J(\alpha + a_i)}$	**Naive**
β_1	3.602	3.602(1.231)	3.602(1.007)	1.590(0.491)	3.590(1.327)	5.086(0.859)
β_2	-1.976	-1.971(1.587)	-1.879(0.483)	-0.022(0.425)	-1.919(0.437)	-0.683(0.382)
β_3	4.326	4.330(1.123)	4.966(0.991)	1.250(0.526)	4.372(0.795)	17.284(0.490)
β_4	0.582	0.589(0.974)	0.702(0.565)	1.041(0.483)	0.735(0.492)	1.878(0.463)
β_5	0.107	0.114(1.142)	0.136(0.582)	0.218(0.487)	0.416(0.548)	1.520(0.502)
β_6	-0.527	-0.518(1.189)	-0.556(0.479)	-0.890(0.433)	-0.540(0.451)	0.601(0.382)
β_7	1.701	1.705(0.868)	1.817(0.643)	1.179(0.499)	1.784(0.686)	3.003(0.588)
β_8	-0.101	-0.092(0.080)	-0.040(0.530)	-0.750(0.473)	-0.141(0.481)	1.165(0.428)
β_9	1.505	1.514(0.940)	1.606(0.878)	0.750(0.498)	1.505(0.646)	2.899(0.617)
β_{10}	1.293	1.298(0.868)	1.345(0.568)	1.165(0.493)	1.340(0.5722)	2.073(0.461)
β_{11}	0.876	0.907(0.847)	0.920(0.493)	1.375(0.470)	1.828(0.438)	2.073(0.461)
β_{12}	-3.525	-3.496(2.139)	-3.578(0.541)	-0.173(0.376)	-3.391(0.466)	-2.308(0.403)
β_{13}	0.602	0.604(1.108)	0.704(0.835)	0.230(0.505)	0.530(0.728)	2.051(0.633)
β_{14}	1.892	1.904(0.939)	2.035(0.746)	1.065(0.491)	1.954(0.671)	3.216(0.654)
β_{15}	0.685	0.708(0.816)	0.732(0.445)	1.664(0.461)	0.552(0.445)	1.743(0.419)
σ^2	10	10.200(3.194)	10.398(3.3.225)	6.729(2.594)	9.865(3.141)	8.397(2.898)

Table 2: E*stimates for the success probabilities obtained from the different models fitted to the data. The data were generated by a joint model with selection probability,* $logit(\rho_{ij}) = \alpha_j + a_i$ *The column True gives the true values, Combined gives the results obtained from the combined model, 'J' refers to the results obtained from a joint model with the logit of the selection probabilities given in brackets and Naive indicates results from the model that disregards the selection process.*

$\boldsymbol{\beta_j}$	**True**	**Combined**	$\boldsymbol{J(\alpha_j + a_i)}$	$\boldsymbol{J(\beta_j + a_i)}$	$\boldsymbol{J(\alpha + a_i)}$	**Naive**
β_1	0.837	0.840	0.863	0.676	0.888	0.933
β_2	0.262	0.290	0.288	0.496	0.262	0.417
β_3	0.881	0.884	0.933	0.640	0.929	1.000
β_4	0.562	0.561	0.587	0.617	0.6008	0.671
β_5	0.510	0.509	0.520	0.524	0.558	0.568
β_6	0.441	0.440	0.437	0.396	0.431	0.811
β_7	0.678	0.679	0.711	0.632	0.727	0.568
β_8	0.487	0.486	0.499	.412	0.484	0.632
β_9	0.659	0.695	0.689	0.585	0.695	0.802
β_{10}	0.637	0.637	0.660	0.631	0.676	0.728
β_{11}	0.593	0.596	0.612	0.654	0.732	0.728
β_{12}	0.166	0.164	0.142	0.478	0.130	0.244
β_{13}	0.568	0.567	0.587	0.525	0.573	0.726
β_{14}	0.700	0.0.703	0.733	0.620	0.746	0.827
β_{15}	0.573	0.574	0.591	0.684	0.576	0.694

Table 3: *Relative bias for the parameters governing the rating process obtained from the different models fitted to the data. The data were generated by a joint model with selection probability, $logit(\rho_{ij}) = \alpha_j + a_i$. The column Combined gives the results obtained from the combined model, 'J' refers to the results obtained from a joint model with the logit of the selection probabilities given in brackets and Naive indicates results from the model that disregards the selection process.*

$\boldsymbol{\beta_j}$	**Combined**	$\boldsymbol{J(\alpha_j + a_i)}$	$\boldsymbol{J(\beta_j + a_i)}$	$\boldsymbol{J(\alpha + a_i)}$	**Naive**
β_1	0.000	0.000	0.559	0.003	0.412
β_2	0.002	0.049	0.989	0.057	0.654
β_3	0.001	0.148	0.711	0.011	2.995
β_4	0.013	0.207	0.788	0.262	2.227
β_5	0.066	0.276	1.042	0.290	1.236
β_6	0.016	0.056	0.689	0.0254	2.140
β_7	0.002	0.068	0.307	0.049	0.765
β_8	0.089	0.600	6.449	0.404	12.564
β_9	0.006	0.067	0.502	0.000	0.926
β_{10}	0.004	0.041	0.099	0.036	0.996
β_{11}	0.035	0.050	0.569	1.087	1.558
β_{12}	0.008	0.053	0.951	0.038	0.345
β_{13}	0.003	0.169	0.618	0.120	2.406
β_{14}	0.007	0.076	0.437	0.033	0.700
β_{15}	0.034	0.068	1.430	0.194	1.547
σ^2	0.022	0.040	3.257	0.014	0.159

The $Relativebias = |(\widehat{\beta_j} - \beta)/\beta|$ where $\widehat{\beta_j}$ are the fixed effects parameter of the j^{th} cluster due to the fitted models and β_j is the fixed effects parameter of the true model.

Table 4: *Confidence interval for the parameters governing the rating process obtained from the different models fitted to the data. The data were generated by a joint model with selection probability, $logit(\rho_{ij}) = \alpha_j + a_i$. The column Combined gives the results obtained from the combined model, 'J' refers to the results obtained from a joint model with the logit of the selection probabilities given in brackets and Naive indicates results from the model that disregards the selection process*

$\boldsymbol{\beta_j}$	**Combined**	$\boldsymbol{J(\alpha_j + a_i)}$	$\boldsymbol{J(\beta_j + a_i)}$	$\boldsymbol{J(\alpha + a_i)}$	**Naive**
β_1	[1.1897;6.0145]	[1.6277;5.5759]	[0.6275;2.5519]	[0.9864;6.1894]	[3.4019;6.7707]
β_2	[-5.0812;1.386]	[-2.8257;6.9061]	[-0.8541;0.8094]	[-2.7687;-1.0541]	[-1.4323;0.0659]
β_3	[2.1301;6.5314]	[3.0261;6.906]	[0.2202;2.2802]	[2.8145;5.9289]	[16.3229;18.2441]
β_4	[-1.3189;2.4977]	[-0.4052;1.8096]	[0.1041;1.9959]	[-0.2293;1.6985]	[0.8946;2.8620]
β_5	[-2.1237;2.3513]	[-1.0046;1.2772]	[-0.7372;1.1734]	[-0.6576;1.4894]	[0.5362;2.5036]
β_6	[-2.8494;1.8126]	[-1.4942;0.3816]	[-1.7393;-0.0407]	[-1.4236;0.3428]	[-0.1483;1.3505]
β_7	[0.0043;3.4053]	[0.5571;3.0773]	[0.2016;2.1558]	[0.4355;3.1321]	[1.8511;4.15541]
β_8	[-0.2477;0.0643]	[-0.9987;1.0793]	[-1.6766;0.1764]	[-1.084;0.8012]	[0.3232;2.0033]
β_9	[-0.3284;3.3568]	[-0.1153;3.3269]	[-0.2161;1.7157]	[0.2397;2.7709]	[1.6904;4.1078]
β_{10}	[-0.4035;3.0003]	[0.2317;2.4591]	[0.1988;2.1306]	[0.2181;2.4611]	[1.5435;3.6183]
β_{11}	[-0.7529;2.5669]	[-0.0456;1.8858]	[0.4533;2.2961]	[0.9754;2.6810]	[1.1687;2.9763]
β_{12}	[-7.6887;0.6961]	[-4.6391;-2.5169]	[-0.9095;0.5645]	[-4.3039;-2.4787]	[-3.0964;1.5186]
β_{13}	[-1.5679;2.7763]	[-0.9331;2.3413]	[-0.7602;1.2210]	[-0.8971;1.9567]	[0.8101;3.2923]
β_{14}	[0.0632;3.7452]	[0.5736;3.4960]	[0.1021;2.0283]	[0.6394;3.2690]	[1.9336;4.4984]
β_{15}	[-0.8923;2.3079]	[-0.1408;1.6040]	[0.7602;2.5678]	[-0.3205;1.4239]	[0.9224;2.5640]

4 Discussion

To summarise the expert opinion, different modelling techniques were used. First Naive model was fitted on the data to check the impact of ignoring selection process. The results it reveal that ignoring the selection process, has a seriously impact on the parameter estimates on comparing with the true values of the model as observed in Table (1) above. This similarly had an impact on the Confidence Intervals(C.I) like narrow $C.I^s$ (Table 4). This can be as a result of ignoring some information from the selection process. Therefore, these results are wrong and should not be relied on for any inferences. The relative bias associated with this model are also are high (Table 3), which confirm the aforementioned statement.

The second approach employed was the joint modelling which assumes conditional absence of selection bias and follows a different path to estimate the parameters of interest and in the process takes into account the selection process. Three joint models were fitted to assess the effect if any misspecification. The first joint model was correctly specified in all setting, the second and third joint models, the mean of the selection process were misspecified to check on the impact of misspecification on the model. The second joint model assumed that the mean of the rating and selection processes being the same. However, each process having different rater and selection specific random effects. Then the third one assumed a constant mean (α) effect of the selection process. Based on the results of the estimates in Table (1), it can be clearly seen that misspecification of the mean had an impact on both the estimates of the clusters effects and variances as it led to reduction in variance. This might mislead inferences as one might rush to use estimates with high precision and yet they are not correct. Misspecification of the mean or ignoring the selection process also had in impact in terms of relative bias as this led to shot-up (Table 3). For example the

second joint model produced relative biases ranging between 30% to 655% which appear to be so high.

Another approach used was of the exponential family suggested by Booth *et al.* (2003) and Molenberghs *et al.,* (2010) to account for the selection process and the results were compared with the true model. Its biases were so small which possibly meant that the combined model accounted for all the information in the dataset. The values of the estimates of the fixed cluster estimates effects were quite close to those of the true values and the correctly specified joint model (Table 1). However the combined model is less precise compared to others models which my be the price to pay for being more robust. In terms of C.I coverages, the combined was also doing well as most percentages of finding a cluster estimate in the 95% confidence coefficient were as high as above 92% (Table 5 of the appendix). Therefore a combined model can be used for sensitivity analysis in this case and in situations where the joint model is not doing well, one can use the results from the combined model for inferences.

If one want to interpret the estimates, the variance $\widehat{\sigma^2}$ represents the inter-expert variability. For the combined model and correctly specified joint model they were ≈ 10 which is almost as the one for the true values. However this values of inter-expert variability seems to be high meaning that there is need to select the experts from a more uniform population in order to reduce this heterogeneity. Both the fixed cluster effects and probabilities can be interpreted at a cluster specific level. The fixed cluster effects β_j can be interpreted in form odds of ratios.

5 Conclusion

In handling expert opinion in the drug discovery process, one would want to ignore the selection process when say analyising the data due to the complexities involved. However, ignoring this selection may seriously threaten the validity of the results. As a consequence, one often needs to jointly model the rating and selection processes in order to avoid bias in estimating the parameters of interest. Research have shown that in even carefully designed study, the bias might still exist. Ideally, one would like to know all the factors influencing the selection process beforehand. But in practice, such information is seldom available and making assumptions on the selection process is almost inescapable, and if these assumptions are wrong, estimates and inferences may be wrong as well. From the above results, it has been noted that misspecifying or ignoring the selection process may introduce severe bias in the estimates of the parameters.

As a matter for sensitivity, we used another model called the combined model. The results showed that, unlike the naive and the incorrectly specified joint model, the correctly specified joint model approach and the combined model seems to produce unbiased. Therefore the combined model can act as alternative model for the joint model.

6 References

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.

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.

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

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**, 279311.

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

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.

http://www.sciencemag.org/site/products/ddbt_0207_Final.xhtml(accessed on 01/02/2013).

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.

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.

Molenberghs, G., Verbeke, G., Demetrio, C., and Vieira, A. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Sci-*

ence 2010; **25**:325-347.

Sheridan R. P., M. D. Miller and D. J. K. S. K. Underwood, "Chemical Similarity Using Geometric Atom Pair Descriptors," *J. Chem. Inf. Comput. Sci.* 1996, **vol. 36**, **pp.** 128-136, .

Paul S. M., D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg and A. L. Schacht, "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nature Reviews Drug Discovery* 2010, **vol. 9**, pp. 203-214, .

Kearsley S. K., S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley and R. P. Sheridan, "Chemical Similarity Using Physiochemical Property Descriptors," *J. Chem. Inf. Comput. Sci.* 1996, **vol. 36, no. 1, pp.** 118-127, .

7 Appendix

Table 5: *Percentage of confidence interval coverage for the parameters governing the rating process obtained from the different models fitted to the data. The data were generated by a joint model with selection probability, $logit(\rho_{ij}) = \alpha_j + a_i$. The column Combined gives the results obtained from the combined model, J refers to the results obtained from a joint model with the logit of the selection probabilities given in brackets and Naive indicates results from the model that disregards the selection process*

$\boldsymbol{\beta_j}$	**Combined**	$\boldsymbol{J(\alpha_j + a_i)}$	$\boldsymbol{J(\beta_j + a_i)}$	$\boldsymbol{J(\alpha + a_i)}$	**Naive**
β_1	96.8	90.2	53.5	89.5	73.5
β_2	100	100	60.5	96.8	95.0
β_3	94.9	96.2	55.5	74.2	53.0
β_4	98.0	100	100	95.9	95.0
β_5	97.4	98.0	46.5	95.5	93.0
β_6	99.5	100	80.5	93.6	95.0
β_7	96.2	98.4	60.5	96.8	88.5
β_8	83.6	97.5	84.1	81.9	94.0
β_9	99.0	96.1	72.0	97.6	90.0
β_{10}	98.0	98.0	46.3	96.8	93.5
β_{11}	98.0	96.1	74.4	93.6	93.5
β_{12}	100.0	16.1	94.9	90.4	97.5
β_{13}	98.5	98.0	73.9	94.0	93.0
β_{14}	97.4	92.2	02.6	91.1	93.0
β_{15}	99.5	96.1	46.1	97.2	94.0

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Modeling of selection process during evaluation of clusters chemical of chemical compounds in the drug discovery process

Richting: **Master of Statistics-Biostatistics**
Jaar: **2014**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Amwonya, David

Datum: **6/02/2014**