

Modelling and simulation SIG update: Best Practice document

Michael O'Kelly, member of Modelling and Simulation SIG



Summary

- Work on Best Practice since 2011.
 - > European Federation of Pharmaceutical Industries and Associations (EFPIA) MID3 working group
 - > EFSPI Special Interest Group (SIG)
- PSI Board agrees to adopt SIG Best Practice proposal pending publication of the proposal in *Pharmaceutical Statistics*.
- EFSPI Modelling and Simulation SIG is working with MID3 to gain agreement among practitioners for Best Practice in modelling and simulation globally.

EMA: “Best practice” depends on importance of project



EMA-EFPIA Modelling
and Simulation Workshop

Good practices and next steps

Robert Hemmings, EMA

M&S good practices

- Different standards for different exercises (L,M,H)
- Standard should be high!
 - Assumptions (not only mathematical)
 - Model building rationale
 - Model testing
 - Inference
 - Sensitivity analyses / Challenge assumptions
 - Reporting
- Detail of regulatory response might be vary according to impact

EMA: “Best practice” depends on importance of project



EMA-EFPIA Modelling
and Simulation Workshop

Good practices and next steps

Robert Hemmings, EMA

M&S good practices

- Different standards for different exercises (L,M,H)
- Standard should be high!
 - Assumptions (not only mathematical)
 - Model building rationale
 - Model testing
 - Inference
 - Sensitivity analyses / Challenge assumptions
 - Reporting
- Detail of regulatory response might be vary according to impact

MID3 paper on Good Practices

- MID3 paper “Good Practices in Model-Informed Drug Discovery and Development (MID3): practice, application and documentation.”
 - › Wide industry representation: Pfizer, Bayer, Roche, AstraZeneca, GSK, J&J, Merck, BI, Novartis, Novo Nordisk.
- Paper plus supplementary spreadsheet of 103 MID3 example applications, published January 2016.

MID3 headings for Good Practice

Components of Good Practice plans		
Analysis plan	Simulation plan	Report
<ul style="list-style-type: none"> • Introduction • Objectives • Data plan • Data exploration • Methods <ul style="list-style-type: none"> • Model building • Selection+evaluation • Qualification • Assumptions • Results 	<ul style="list-style-type: none"> • Introduction • Objectives • Additional data • Methods <ul style="list-style-type: none"> • Identify model • Limitations • Qualification • Assumptions • Results 	<ul style="list-style-type: none"> • Synopsis • Introduction • Objectives • Data • Methods <ul style="list-style-type: none"> • Identify model • Limitations • Qualification • Assumptions • Results • Applications/simulations • Discussion • Conclusion • Appendices

MID3 headings for Good Practice

Components of Good Practice plans		
Analysis plan	Simulation plan	Report
<ul style="list-style-type: none"> • Introduction • Objectives • Data plan • Data exploration • Methods <ul style="list-style-type: none"> • Model building • Selection+evaluation • Qualification • Assumptions • Results 	<ul style="list-style-type: none"> • Introduction • Objectives • Additional data • Methods <ul style="list-style-type: none"> • Identify model • Limitations • Qualification • Assumptions • Results 	<ul style="list-style-type: none"> • Synopsis • Introduction • Objectives • Data • Methods <ul style="list-style-type: none"> • Identify model • Limitations • Qualification • Assumptions • Results • Applications/simulations • Discussion • Conclusion • Appendices

MID3 headings for Good Practice – includes recommendations for each heading

Components of Good Practice plans		
Analysis plan	Simulation plan	Report
<ul style="list-style-type: none"> • Introduction • Objectives • Data plan • Data exploration • Methods <ul style="list-style-type: none"> • Model building • Selection+evaluation • Qualification • Assumptions • Results 	<ul style="list-style-type: none"> • Introduction • Objectives • Additional data • Methods <ul style="list-style-type: none"> • Identify model • Limitations • Qualification • Assumptions • Results 	<ul style="list-style-type: none"> • Synopsis • Introduction • Objectives • Data • Methods <ul style="list-style-type: none"> • Identify model • Limitations • Qualification • Assumptions • Results • Applications/simulations • Discussion • Conclusion • Appendices

EFSPI proposed Best Practice document

- Authored by volunteers from the SIG
 - › SIG members from variety of pharmaceutical companies.
 - › All authors of the Best Practice document from contract research organizations.
 - › Agreed to be adopted by Board of PSI.
- PSI Board agrees to adopt SIG Best Practice proposal pending its publication in *Pharmaceutical Statistics*.
 - › Best Practice paper submitted to *Pharmaceutical Statistics* November 2016.
 - › Paper is currently in review (second round of referees' comments).

EFSPI SIG Best Practice recommendations – during, after, and next time around...

- Template for specification
 - › describe listed key elements or justify why not
 - › justify level of detail of the pre-specification.
- Quality control – level of QC should be appropriate -
 - › from simple review of specification (low-impact project)
 - › to independent programming of project (some high-impact projects)
- Presentation of results
 - › may vary depending on audience – plan in advance outputs for each audience
 - › Statistical clarity: use of confidence intervals; operating characteristics; measure of stochastic variability in simulations....
- Changes to specification
 - › Specification should be auditable, e.g.,
 - » revision history
 - » formal amendment (as in protocol amendment)
 - » include old versions as appendices

Principle: do what is necessary for Best Practice, but not more

- SIG document allows the flexibility necessary for Best Practice in this area where the regulatory and scientific importance of the projects varies widely.

Example best-practice specification for low-impact work

Planning and Reporting for Projects that Involve Modelling and Simulation Best Practice Document

Appendix B: example specification with a low level of detail

Using simulated data to verify an estimate of probability of success

Specification of simulations

B.1 Introduction

Given five treatment development programs with known probability of success, it is desired to know the probability of zero successes and of four and five successes. These probabilities have been calculated analytically. It is requested that a simulation be run to verify that the analysis is correct.

Since this is a one-off query on whose evidence alone no decision will be made, this is judged a project of low importance. Therefore the clinical background is not described; nor are metrics and criteria for decisions appropriate.

B.2 Simulation and analysis/design

As noted, this project is of low importance and no decision will be made by it alone. Therefore the description of the elements of the simulation and analysis will not be detailed and some elements are not applicable.

B.2.1 Scenarios assumed and assumptions made

Probabilities of 0.1, 0.2, 0.2, 0.05 and 0.4 were given for programs 1-5, respectively. Since the objective was simple verification of an existing calculation, no justification is given here of these probabilities. Since the question answered is theoretical, just one given scenario is used.

B.2.1.1 Sensitivity analyses

This project is not required to assess assumptions, so sensitivity to assumptions is not planned to be analysed via sensitivity analyses

B.2.2 Data sets generated

Temporary sets of binary outcomes will be generated. Data will not be bootstrapped because a simple verification is sufficient. Three million binary outcomes are simulated for each program.

Page 1

B.2.3 Statistical analysis

The number of instances of zero, four and five successes was calculated for each of the 3 million simulations, and the probability of zero, four and five successes in a simulated instance was calculated and plotted.

B.2.4 Operating characteristics

Given that this modelling and simulation task is to be a sanity check, the number of simulations required to achieve a given accuracy with 5% confidence will be approximated. The probability of five successes is small ($<1/100$) so a precision of 0.001 is desired. Using the formula of Burton *et al.* (2006), with $\alpha=0.05$, and approximating the variance of the probabilities as $5 * p(1-p)$ where $p=0.2$, 3 million simulations will provide precision of approximately 0.001.

B.2.5 Logistics

The R language package `myBinaryEP` will be used to simulate the binary outcomes. The package allows for correlations between the outcomes, but this was not required for the primary objective. R version 3.0.1 will be used. See Appendix for the R code used. The seed used was 1.

B.3 Quality control

Given that this modelling and simulation task is of low importance and will not of itself lead to a decision, the specification will be submitted to the requestor of the calculation, but no further QC of the production of results is planned. The output will be checked against the requestor's calculations.

B.4 Presentation of results

A table will be presented of the probability of zero, four and five successes among the five programs, calculated as the proportion of instances of zero, four and five successes in 3 million simulations of the five programs. The number of instances will also be plotted in a histogram with one stack for each level of successes, a stack for zero successes, 1 success, ... and so on up to five successes. These outputs are judged sufficient to act as a check of the analytic estimates, which is the objective of this project.

The precision of the result (standard error) will be presented in a footnote to the plot. Given the inclusion of precision, no confidence intervals will be presented. Given the theoretical nature of the problem and the corresponding simplicity of the simulation, no bias is to be expected in the simulation-based estimates.

The results of the modelling and simulation will not be stored. The R code will be stored in [location]. A note of the contents of the table output will be included as a comment in the R code.

Page 2

Example best-practice specification, high-impact work

Planning and Reporting for Projects that Involve Modelling and Simulation Best Practice Document

Appendix A: example specification with a medium-high level of detail

Using simulated data to assess analyses of negative binomial outcomes with missing data

Specification of simulations

A.1 Summary

The objective of the project is to produce a user-friendly means that has been tested using a combination of real and simulated recurrent event data. This means **eligibility** was a Multiple Deprivation (MD) sensitivity analysis method to implement a variety of assumptions, including reference-based assumptions, for missing data, in order to assess accuracy involving recurrent event outcomes that are missing not at random (MNAR), with an approach that assesses the negative binomial distribution. This means assessing the treatment effect of a number of different treatments, compared against a control or reference group, using difference of time square means.

A.2 Introduction of the specification

This document has been modified on the best practice document of the Special Interest Group for Modelling and Simulation of RCT [1]. The objective of the simulation exercise is to assess, with respect to the MD sensitivity analysis:

- How the Type I error rate under the null hypothesis of no treatment difference
- Power to detect a treatment difference when one exists

These objectives will be addressed by simulating datasets where (1) the treatment effects of two different treatment types are the same, (2) and the treatment effects of two different treatment types are different. The simulation will also determine the impact of percentage dropout on both the power and the Type I error rate.

Page 1

A.3 Simulation and analysis design

The objective of the assessment of type I error and power is to help to validate the means as an implementation of an analysis of recurrent event data under the missing-at-random (MAR) assumption, where the distribution of the outcome is negative binomial. To validate the implementation, results from the MD approach will be compared with those of a direct likelihood one, using the standard approach. The standard approach for negative binomial outcomes is a direct likelihood analysis, using a generalized linear model, with the varying rates of response accounted for by an offset term that is equal to the log of the time elapsed. If the MD approach correctly implements the analysis for outcomes that are distributed as negative binomial under MAR, results from the standard approach and the MD approach should agree. Therefore, results in this section from MD analyses of the simulated datasets agree with the standard analysis. The MD approach can be regarded as valid and appropriate. This modelling-by-difference outcome (read correlation), is a significant result, to be discussed as to whether to use the implementation in regular clinical trials. Therefore, the project is judged to be of medium importance. There will be a moderate amount of time given in the following sections on Simulation and Analysis to ensure that no evidence is reproduced.

A.3.1 Scenario assessed and assumptions made

As mentioned in Section A.2, there are two main objectives for the simulation exercise; the datasets will be simulated in two different ways in order to fulfil these objectives. Scenario one will address the Type I error rate of the method, as implemented by the means, and scenario two will address the power of the method, as implemented by the means. To do this, recurrent event datasets will be simulated in two treatment groups having the same outcomes with the rate of recurrent events for scenario one and differing with respect to that association for scenario two. In addition to treatment group, other approaches will be included, such as being compared, with two comparisons, and two comparisons separately. An underlying subject-specific effect will be included, using random sampling from a gamma distribution. Datasets are simulated using this method in order to simulate the negative binomial, as described by Kuznetsov et al. [2].

Meaning data is generated by simulating a rate of dropout that tends to increase as an event or occurrence occurs. Both recurrent events and dropouts are simulated, independently from each other, for each time point, using a Bernoulli distribution. The probability of an event is modelled as a linear combination (1) of each subject's baseline covariates, which has been **quasi-randomly** multiplied by the subject-specific effect (2), and then the previous values being converted to a probability value using the **aggregated back-transformation** of $y = \frac{e^x}{1+e^x}$, which is described in Kuznetsov et al. [2]. The probability of dropout is a linear combination of each subject's baseline covariates (3) which will be generated, given that the subject is not censored, multiplied by the subject-specific effect (4), and then the previous values being converted to a probability value using the **logit back-transformation** of $p = \frac{e^x}{1+e^x}$.

We are not currently addressing the model's ability to account for multiple different types of dropouts (e.g. death, adverse event, subject dropout); therefore, there will only be two event types

Page 2

in the response variable, recurrent event (1) and dropout (2). Calculations based on all simulated datasets, using the highlighting means, will be performed using the following: An Random (GAR) assumption, without any delta adjustments of input values.

The response coefficients for the models to be used in all datasets when generating the recurrent events will be based on real data from the bladder tumour recurrence dataset, taken from the R package 'survival' and described in [3] and [4]. The true values for the simulated continuous variables are set to equal to the true values for the continuous variable number and after divided by the random values. When the treatment effects are set to be equal, (the null case), the same base values are used for both of the simulated groups. When the treatment effect is not to be different (the alternative case), event analysis will be performed in order to test the model's sensitivity to the degree of difference in true values. The difference in true values for treatment effect will be set to 0.45, 0.05 and 0.15, generating **expected** treatment differences of 0.1, 0.47 and 0.7.

A.3.1.1 Sensitivity Analysis

A sensitivity analysis will be performed on both scenarios, to assess the sensitivity of the means to change in the percentage **dropout** of the data. This will be performed by altering the true values that generate the probability of dropout to produce simulated datasets where an approximate percentage of subjects have dropped out of the study early. Seven rates of dropout will be compared, 1%, 5%, 10%, 15%, 20%, 25% and 40%. The number of simulations during a significant difference in the treatment effects will be counted, for both the null and alternative scenarios, to determine the effect of percentage dropout on both the type I error rate and the power of the means. As noted, we will also test the sensitivity of the means to changes in the percentage of dropout in the data when increasing the power and the type I error rate. The true values used to determine the rate of dropout will be randomly calibrated to produce seven rates of simulation, comparing seven different percentage of subject dropout: 1%, 5%, 10%, 15%, 20%, 25% and 40% dropout.

A.3.1.2 Scenario generated

The simulated datasets will be modelled in a normal form, which allows for multiple models per event, and whose observations correspond to a subject over time. One set of seven simulated datasets will be generated for the three scenarios to test the type I error rate. When testing the power, three sets of simulations will be generated. Some simulations will be generated for each set to be tested using the **aggregated** back-transformation described in Section A.2. These datasets will include a continuous random variable to identify each simulation, a subject ID variable to identify each subject, an event variable to determine

Page 3

whether the observation is a recurrent event or a dropout, a variable for each baseline covariate (two continuous and two categorical, including treatment) and a time variable to identify the time point at which each event or dropout occurred. Data will not be homogenized because we are using specific scenarios and we want to control those scenarios artificially; however, as noted, the true coefficients were based upon results on values found in real data.

A.3.1.3 Statistical analysis

For scenario one, the proportion of times the null case gives a false positive result is calculated. The MD result is then compared to the same result using the standard Direct Likelihood approach in order to get a comparison in performance between the two methods. For scenario two, the alternative hypothesis, the two analyses are compared on the datasets where the treatment effects are different and the power is calculated and compared. The sensitivity analysis will test the degree to which the type I error and power of the means are dependent on percentage dropout. The proportion of false positive and false negative results will be calculated. Again, these MD results are compared in performance against the Direct Likelihood analysis results.

A.3.4 Quantifying characteristics

The number of simulations required to give an accurate measure and/or to achieve a comparison with sufficient accuracy, with a specified confidence level and clearance to the true or desired value, can be calculated using the formula from Burton et al. [5]. This calculation estimates the required number of simulations based on the accuracy of six scenarios of interest. The number of simulations required (N) is calculated as:

$$N = \frac{z_{1-\alpha/2}^2}{\delta^2} \left(\frac{p(1-p)}{p} \right)$$

The parameter δ is the specified level of accuracy, or the permitted difference from the true or desired value, α represents the simulated deviation, $z_{1-\alpha/2}$ is the specified quantile of the standard normal distribution and p is the significance level required.

(Details of the calculations have been omitted for this example)

A.3.5 Logistics

SAS v9.4 has been used for the primary software support. The negative binomial was approximated via the Bernoulli distribution with a gamma random effect. Events and dropouts are simulated by sampling from Bernoulli distributions, where the probability of an event or dropout is, in turn, by employing the baseline covariates by their corresponding true values. These values are simulated and aggregated, when it is assumed that the log of the number of events has a linear relationship with the predictor. Then the aggregated transformation is used to calculate the probability of an event (1) or dropout (2) given the covariates. The probability of dropout is also based based on the number of

Page 4

concurrent events occur. Kuznetsov et al. (2014) describe how the negative binomial can be generated by using the Poisson distribution mixed with gamma. The explanation for further details of the Poisson process using a Weibull distribution.

The code to be used when generating from each distribution (using the rand function) is 22 to begin with but this is recommended by 1 at the start of each simulation. Using a different code for the generation of each dataset allows there to be completely independent of each other.

A.4 Quality control

The specification was independently reviewed. The core functions of the code will be unit tested to verify the operation/behavior. The code is planned to be submitted to the CMA special working group with page in mcgill@mcgill.ca. If the project is included as of high importance, the delivery timeline should be added. The project results will be quality controlled via independent programming.

A.5 Presentation of results

Tables will be generated for both scenarios. For all simulation analyses, in each of the two tables the following columns will be included: the percentage coverage, the percentage of significant differences using the MD method, the percentage coverage using the MD method, the percentage coverage using the direct likelihood method and the variance ratio. The variance ratio will be the ratio of the observed variance to Rubin's estimate of variance. The output from each means analysis, which provides the LOG-RATE and difference of LOG-RATE, will be stored for an ease of examination.

A.6 References

1. O'Keefe M, Anonim V, Campbell C, Shenkin S. Proposed Best Practice for Projects that Involve Modelling and Simulation. *Pharmaceutical Statistics* (submitted).
2. Kuznetsov O, Rogov A, Mordukhai M, Kuznetsov M (2014) Using delta sensitivity analysis for recurrent event data using generalized regression. *Pharmaceutical Statistics* 13, 4, 218-234.
3. Anderson DR, Hartzberg AZL (1982) *Delta: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag, New York.
4. Witt U, Lin DY, D'Amico AP (1998) Empirical analysis of recurrences: resampling-based tests. *Statistica Sinica* 18, 1, 1-14.
5. Barone A, Arora DC, Rayson P, Boller SL (2000) *Using Design of Simulation Studies in Medical Research*. *Statistics in Medicine* 20(4): 237-257-272.
6. Lawton H (1982) *Stochastic and probability theory and statistical inference*. John Wiley & Sons, Chichester, 1982.

Page 5

A.7 Supplement

Justification of the use of Bernoulli and Gamma distributions when generating Poisson event data and the Negative Binomial distribution.

Assume that the recurrent event data in time interval T are generated by a Poisson process with parameter λ . How we describe how this data can be approximated by using discrete event simulation with Bernoulli trials. Consider a sufficiently short time Δt , such that probability close to one only occurs either 0 or 1 event. Indeed for small Δt the probability of one event occurring is termed 'small' in proportion to the length of the interval Δt . Using the properties of a Poisson process we get that the probability to have more than one event within Δt is the probability of an event is approximately $\lambda \Delta t + \frac{1}{2} \lambda^2 \Delta t^2 + \dots$. Then the events in the non-overlapping intervals of this length, Δt , can be considered to be independent Bernoulli trials.

Now we can describe only zero, 1, two, $n = 1$ or n non-overlapping intervals of equal length. These independent Bernoulli trials each have a probability of an event $p = \lambda \Delta t$. Considering Bernoulli trials with probability p we get that the number of events in the study of length t is binomial with parameters (n, p) , where

$$P_n(k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} (\lambda \Delta t)^k (1 - \lambda \Delta t)^{n-k}$$

It is well-known [6, 7] that for a large n and small probability p such that $n p \rightarrow \lambda$, the binomial distribution is approximated by a Poisson distribution with parameter λ .

Thus,

$$P_n(k) \approx \frac{e^{-\lambda} \lambda^k}{k!}$$

So each $\Delta t = 0.1, 1$. This proves that when events generated that are consistent with the above assumptions, the number of events, X , in an interval of fixed length, t , has a probability function of

$$P_n(k) = \frac{e^{-\lambda} \lambda^k}{k!}, \lambda = 0.3, 1, \dots$$

where X is the Poisson random variable with parameter λ . This shows how the Bernoulli trials can be used to approximate a Poisson process in continuous time.

For recurrent events the negative binomial distribution can be derived as a mixture of Poisson distributions where the resulting distribution are mixed jointly on the intensity of the Poisson process. The mixing distribution is a gamma distribution with mean λ . The negative binomial distribution is also called a gamma-Poisson distribution for this reason. The gamma-Poisson

Page 6

model is a special case of generalized linear mixed models with a Poisson distribution and log link. In this model, observations Y_i are made on the i th subject in scenario 1:

$$Y_i | \eta_i \sim \text{Poisson}(\eta_i)$$

and

$$\ln(\eta_i) = X_i \beta + \eta_i$$

where η_i is a subject specific effect with some distribution on the real line. The design matrix X_i accounts for both the treatment type and the baseline covariates. $\text{Cov}(\eta_i) = \text{diag}(\sigma^2)$, $\eta_i \sim \text{exp}(\lambda)$, then

$$Y_i | \eta_i \sim \text{Poisson}(\lambda e^{X_i \beta + \eta_i})$$

The negative binomial model is a special case of this generalized linear mixed model with a Poisson distribution and log link, where η_i has a gamma distribution.

In the planned simulations, this is generated from a gamma distribution and multiplied by the inverse mean, λ , which is the exponential linear predictor described in Section A.2.1.

Page 7

Best Practice for modelling and simulation, the work of the two groups, MID3 and EFSPI

- Agreement all aspects of the process
- Some differences in emphasis
- The two groups are working together to promote good practice
- The two groups participated at session on Best Practice at 2016 annual PSI conference, Berlin.
- The two groups will participate along with FDA presenters in 2016 September ASA-Biopharmaceutical Section workshop in Washington.

Summary

EFSPI Best Practice document can be used as a tool or template to implement Best Practice as described by MID3 and/or EFSPI.

MID3 and EFSPI SIG share vision of good practice harmonised across the uses of modelling and simulation.

Questions?

Back-up slides

Best Practice in modelling and simulation

MID3

- When to use simulation
- Key elements for a good plan
- Quality control
- Iterative nature of the MID3 process

EFSPI

- When to use simulation
- Key elements for a good plan
- Quality control
- Iterative nature of modelling and simulation

Best Practice in modelling and simulation

MID3

- Agreed across 10 companies.
- Emphasis on integrating MID3 into the general pharmaceutical development process
- Three planning documents.
- Lists key recommended elements.
- Report: specifies sections, with potentially different audiences.

EFSPI

- Authored by SIG, to be adopted by PSI.
- Emphasis on providing a tool for Best Practice for the working statistician
- One specification for a project.
- Emphasis on flexibility – specification should include key elements or justify their absence.

Best Practice in modelling and simulation

MID3

- Emphasis on hypothesis **generation** rather than hypothesis **testing**.
- Tends not to go into detail on technical requirements.

EFSPI

- Allows for possibility of hypothesis testing.
- Suggests including “less likely” scenarios in simulations.
- Considers technical detail, e.g., operating characteristics; use of confidence intervals; measure of stochastic variability in simulations; randomisation seed; software version.

Best Practice in modelling and simulation

MID3

- Emphasis on hypothesis **generation** rather than hypothesis **testing**.
- Tends not to go into detail on technical requirements.

EFSPI

- Allows for possibility of hypothesis testing.
- Suggests including “less likely” scenarios in simulations.
- Considers technical detail, e.g., operating characteristics; use of confidence intervals; measure of stochastic variability in simulations; randomisation seed; software version.

EFSPI Best Practice document can be used as a tool or template to implement Best Practice as described by MID3 and/or EFSPI.

MID3 and EFSPI SIG share vision of good practice harmonised across the uses of modelling and simulation.