A measure of stability as a criterion for the verification and analysis of simulation models

Toni Monleón-Getino[a,*], Maria del Carmen Ruiz de Villa[a], Jordi Ocaña Rebull[a]

(a) Department of Statistics, University of Barcelona, Avda. Diagonal 645, 08028 Barcelona, Spain

Abstract

The aim of this study is to define a new statistic, *PVL,* based on the relative distance between the likelihood associated with the simulation replications and the likelihood of the conceptual model. Our results coming from several simulation experiments of a clinical trial show that the *PVL* statistic range can be a good measure of stability to establish when a computational model verifies the underlying conceptual model. *PVL* improves also the analysis of simulation replications because only one statistic is associated with all the simulation replications. As well it presents several verification scenarios, obtained by altering the simulation model, that show the usefulness of *PVL*. Further simulation experiments suggest that a 0 to 20 % range may define adequate limits for the verification problem, if considered from the viewpoint of an equivalence test.

Keywords: Verification; simulation model; stability of a simulation; analysis of replications; simulation computer software; clinical trial; linear mixed model

* **Corresponding author.**

University of Barcelona, Avda. Diagonal 465, 08028 Barcelona, Spain. Tel.: (+34) 93 4037085, fax (+34) 93 934 111 733 , e-mail: amonleong@ub.edu.

1. Introduction

A very important problem in simulation is the reliability of the conceptual and computational models, that is, the validation and verification problem (See for example Sargent (2003) and Nance and Sargent (2002) for an exhaustive summary). The conceptual model is defined as a mathematical model that expresses reality and the computational model (simulation model) as the implementation of a conceptual model via an algorithm or software.

Holford *et al* (2000) indicates that clinical trial simulation is mainly performed for two purposes: to extrapolate experimental results to different conditions from those used to build the conceptual model (time, doses, etc.) and to optimize, considering alternative designs, sample sizes, sampling schemes, etc. The importance of simulation in the field of clinical trials gave rise to a best practices document (Holford, 1999) offering directives, which are accurate enough, but neither specify how to perform the simulations nor how to analyze the results.

The most common experimental designs in this field are repeated measurements and one of the best ways to model them is using mixed models (see e.g. Hand *et al*, 1987). These allow the fixed and the random parts to be built in separately from the effects of the experimental factors, such as time, pharmacological treatment, and covariates such as sex, weight, complementary physiological measures, etc.

Mixed models successfully model correlations between measurements from the same individual at different times (repeated measurements) using a suitable covariance structure. Appendices A and B outline the introductory theory of mixed models, Monte Carlo simulation, and the calculation of the number of simulation replications.

One of the most important problems in clinical trial simulations is how to measure the stability of the simulations (in the sense of the degree to which simulation software can be run without crashing, errors in the coding of the conceptual model or otherwise malfunctioning). From a statistical point of view, validation and verification are goodness of fit (GOF) problems. In this context, the use of statistical hypotheses has many methodological disadvantages, especially for determining the range giving an acceptable precision in the verification of the model. The simulation model is obtained by implementing the model on the specified computer system, which includes programming the conceptual model whose specifications are contained in the simulation model specification. Inferences about the system are obtained by conducting computer experiments (experimenting) on the simulation model (Sargent, 2003).

The process of validation concerns how the conceptual model fits reality and is usually defined to mean "substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model" (Schlesinger *et al*, 1979) . A model should be developed for a specific purpose (or application) and its validity determined with respect to that purpose. If the purpose of a model is to answer a variety of questions, the validity of the model needs to be determined with respect to each question. Operational validation checks that the data generated by the computational model have statistical properties comparable to those of real data. During verification is checked whether the behaviour of the computational model adequately implements the conceptual model, that has been judged adequated before for the purpose or its utility.

Model verification is often defined as "ensuring that the computer programming and implementation of the conceptual model are correct" and is the definition is internationally accepted (Sargent, 2003; Balci, 1998) and adopted here. The most recent tendencies in simulation (Thacker *et al*, 2005) concludes that verification is concerned with identifying and removing errors in the model by comparing numerical solutions to analytical or highly accurate benchmark solutions. Validation, on the other hand, is concerned with quantifying the accuracy of the model by comparing numerical solutions to experimental data. In short, verification deals with the mathematics associated with the model, whereas validation deals with the physics associated with the model (Roach, 1998). Because mathematical errors can cancel out, giving the impression of correctness (right answer for the wrong reason), verification should be performed before the validation activity begins .

Sargent (2003) and Nance *et al* (2002) described four different approaches to deciding model validity; two different paradigms that relate verification and validation to the model development process and various methods of model verification techniques, also an extensive described of validation and verification techniques were wide described in Monleón (2005). Otherwise Balci (1998) and Whiner *et al*. (1998) present guidelines for conducting verification, validation, and accreditation (VV&A) of model and simulation (M&S) applications. In the real world of modelling there is a process of model building that involves trying different conceptual models. The best of these models e.g. based on a likelihood ratio test (LRT), is then chosen as the final model to describe the data.

The aim of this study is to define a new statistic, *PVL,* based on a *GOF* criterion, to measure the stability of a simulation or computational model compared to its underlying conceptual model. It is based on the relative difference between the likelihood associated with the simulation replications and the likelihood of the conceptual model. Its use as a verification measure will be illustrated by means of a real clinical trial simulation where the conceptual model is a linear mixed model.

2. A measure of stability for verifying simulation models

To verify computational models compared to the underlying conceptual models using a set of simulation replication results, confidence intervals for the fixed part and variances and covariances for the random part parameters of the conceptual model are currently used (Sargent, 2003); Monleón, 2005; Monleón and Ocaña, 2006; Rodríguez-Barrios *et al*, 2008). Additionally the errors and the statistical significances (p-values) for the main factors (treatment, time, and interaction time × treatments in a typical clinical trial) can also be used. This method has the inconvenience of being extremely demanding in time and it was rejected as a general approach. Most of the techniques described in the literature has been used for validating and verifying the submodels and overall model, and can be used either subjectively or objectively. By "objectively"we mean using some type of statistical test or mathematical procedure, e.g., hypothesis tests or confidence intervals. A combination of techniques is generally used (Sargent, 2003) during V&V process.

Parameters and p-values can show whether the computational model is reasonably comparable to the conceptual model, but the use of statistical hypotheses has many methodological disadvantages. They do not allow us to quantitatively establish the difference that exists between conceptual and computational models. This includes the totality of the simulation replications and obtaining a measurement of its stability when faced with different model disturbances, such as the supposition of new scenarios (variability and design), mistakes during the transcription of the conceptual model in algorithm form, or bugs in the random generation algorithm during simulation. We need a measure to determine the range of the acceptable precision in the computational model verification for its future re-use when the modeller needs to extend the use of the computational model to a new experimental situation.

To solve this problem from a statistical point of view, and considering validation and verification as a GOF problem, we propose the use of a new statistic as a measure

of stability of the simulation model based on a simulated *Log–Likelihood* GOF criterion (*LLK)* calculated as -*2 Res log L*, where *Res log L* is the logarithm of the maximum restricted likelihood function (Monleón and Ocaña, 2006); Monleón *et al,* 2004). This index may be computed in conjunction with other likelihood criteria such as *AIC*, *AICC (*Small variation of *AIC)* and *BIC*, and calculated in every simulation replication.

The likelihood variation percentage compared to the conceptual model likelihood (*PVL, Percentage variation of likelihood*) for every simulation replication (Figure 3) may be defined as:

$$PVL_i = 100\left(\frac{n_a}{n_b}\right)\left|\frac{LLR_b - LLR_i}{LLR_b}\right| \quad (1)$$

where LLR_b is the *LLK* calculated in the conceptual model, LLR_i is the value of *LLK* obtained in simulation replication *i=1 to R*, n_a is the number of cases computed in the conceptual model, n_b is the number of cases simulated in the computational model ($n_b = n_a$ when the sample size and design of the original experiment is used for the simulation). PVL_i is calculated using a GOF based on *LLK* criteria. *AIC*, *AICC* and *BIC* were calculated in the same way. PVL_i is a positive definite distance.

The stability distance between the computational and conceptual models is defined as the *PVL* range (*ul*-*ll*) and composed of the *PVL* upper limit (*ul*) = $PVL_{ul} = Max(PVL_1,\ldots,PVL_R)$ and the *PVL* lower limit (*ll*) $= PVL_{ll} = Min(PVL_1,\ldots,PVL_R)$.

Another measure to verify simulation models could be the *LLR* intra-replications (*PVLIR*) calculated as,

$$PVLIR_i = 100\left|\frac{LLR_2 - LLR_i}{LLR_2}\right| \quad (2)$$

Where LLR_2 is the $Max(LLR_1,\ldots,LLR_R)$ for all *R* estimated in the simulated model. LLR_i corresponds to every value of *LLR* obtained in one of the 1,….,*R* replications.

The next section develops a case study; modelling, simulation, and the results of verification using the proposed measure of stability.

3. Simulation study

Geaghan (2002) described an example of a clinical trial designed to study the impairment that rheumatoid arthritis produces in the mobility of sufferers. Their original

data came from Littell (1996). The trial tested the efficacy of a new drug compared to a standard treatment and a placebo. The treatments were administered over 14 weeks to 57 patients (20 placebo/control (c), 16 standard treatment (s) and 21 new treatment (n)). The trial used repeated measurements performed every two weeks. The end point variable was a mobility scale, with a continuous range from 0 to 100 %. The results are shown in Figure 1.

The authors of the study used a linear mixed model to fit the mobility results and discussed its validity by means of comparing the information produced by the model to the real clinical data.

When fitting such a model, multiple options are available for estimating parameters. The relevance of these options in clinical trial modelling is discussed in Monleón (2005). To cope with possible residual dependencies, one possibility is to use an unstructured covariance matrix, as suggested by Geaghan (2002). This is the option that requires the most parameters to be estimated. Assuming that covariance matrices are diagonal would make it necessary to estimate fewer parameters (Jenrich *et al*, 1986). The impact of incorrect assumptions on the structure of covariance and residual distribution of non-linear mixed models for repeated measurements was studied by El-Halimi *et al* (2004) .

The first-order linear mixed model fitted by Geaghan (2006) to establish the relationship between patient mobility and time, with treatment as an additional factor was:

$$Y_{ij} = (\beta_0+u_{i0}) + (\beta_1+u_{i1})Time_{ij} + e_{ij} \qquad (3)$$

where $j = 1, \ldots, m$ is the index of the measurements through time, $i = 1,\ldots, n$ corresponds to patients or cases, $\boldsymbol{\beta} = (\beta_0, \beta_1)$' are the fixed effects and $\boldsymbol{u_i} = (u_{i0}, u_{i1})$' are the random effects.

Patient mobility is explained by a set of fixed effects (parameter average values) plus random effects associated with deviations of each from the mean value. In the first approximation, the dimension of u is 2, that is, there are as many random effects as parameters in the model (3), and the number of times u is performed equals the number of patients.

In addition, if the model has q experimental factors (drug treatments, interaction between time and treatment, other covariates, etc.) with the possibility of each one having a random term, the equation would be :

$$Y_{ij} = (\beta_0+u_{i0}) + (\beta_1+u_{i1})x_{ij1} +....+ (\beta_q+u_{iq})x_{ijq}+ e_{ij} \quad (4)$$

where Y_{ij} = clinical response variable, j = time, i = patient or case, $\beta_{0...}\beta_q$ are fixed effects, $u_{i0...}$ u_{iq} are random effects, e_{ij} = global error, x_{ijz} = additional covariates, z=1,…, q.

The general model given by (4) is the basis of what is usually called the "conceptual model". Assuming that the model has only three terms (treatment, time, and their interaction) and that no random component is found for the treatment and the interaction, the final conceptual model will be:

$$y_{ij} = (\beta_0+u_{i0}) + \beta_K x_{ijK} + (\beta_2+u_{i2})t_{ij} + \beta_3 t_{ij}\, x_{ijk} + e_{ij} \quad (5)$$

where y_{ij} is the percentage of mobility in case (i), at time t (j), for treatment x(k). It is no problem that the response is measured as a percentage since the right hand side of (5) is Gaussian. A transformed response, such as the logit transformation, could also be considered.

From the dataset described at the beginning of this section, the model parameters were estimated using the algorithm PROC MIXED implemented in the SAS software (SAS, 1992; PROC MIXED) giving the following estimates: u_{i0} follows a N(0, 3.0276 *[2.5291,3.7725]*) distribution. Similarly, $u_{i2} \sim N$(0, 0.1740 *[0.1422, 0.2241]*, β_2 = 0.165 *[0.08205;0.2479]* represent the fixed effect coefficients of the visit times, and the c*ovariance* between u_{i0} and u_{i2} is -0.08889*[-0.2498, 0.07207]*. The residual distribution is $e_{ij} \sim N$(0, 0.7953 *[0.7350,0.8663]*). The values of the treatment effect β_k (-0.8755 *[-2.7827,1.0317]*; -0.9630 *[-2.9887,1.0626]*; 0) and the interaction effect β_3 (-0.1944 *[-0.3110,-0.07788]*; -0.06006 *[-0.1838;-0.06373]*; 0) are the estimates of the differences from basal treatment (n). The 95% confidence intervals (from asymptotic normal theory) for the estimated parameters are indicated in brackets and italics. These confidence intervals should be included to highlight the degree of uncertainty and as the basis for analysing the sensitivity of the computational model by changing the parameter values to other values within these confidence intervals.

The simulation of model (5) may be helpful in investigating new scenarios, such as variation in the power (*1-β*) when decreasing the sample size of the unexpected response to a treatment, or an increase in variability. Data collected in these scenarios

may help optimize future clinical trials related to this type of drug, minimizing the probability of very expensive errors.

Equation (5) can be used as a data-generation mechanism. Transformed into SAS code, the simulation was performed the number of replications required to reach a "previously set accuracy" (Guasch *et al*, 2002). In Appendix B we suggest a formula for calculating the number of simulation replications necessary for a given precision (model acceptability); we consider a 10% precision and 15.8% as a semi-interval confidence interval of the response variable. Holford *et al* (1999) specify that the number of replications must be based on the target precision of the study. To analyze every simulated replication, the same statistical analyses were used as in the original clinical trials, obtaining values for GOF, model parameters, and inferential procedures to check the differences between factor effects, which we discuss in the next section.

4. Application to the verification of the simulation models

We studied confidence intervals for the fixed part, and the variances and covariances for the random part parameters of the conceptual model in order to verify the computational model. We also calculated the error and the statistical significances (p-values) for the main factors (treatment, time, and time/treatment interaction) in this model. Our results show that few model parameters differ in the replications of the computational model (or fall outside the confidence interval) compared to the estimated conceptual model. As we noted above, this method was extremely demanding in terms of the time required.

We further calculated p-values for the drug treatment factor after 100 simulation replications (Table 1, Experiment 1). The simulation results reveal that the p-values for drug treatments agree with expected values when the null hypothesis is true. 47% of experiments had a drug factor p-value < 0.05 (conceptual model p-value for drug treatment < 0.05) and 82% of experiments had a time*drug interaction p-value < 0.05 (conceptual model p-value for drug treatment < 0.05). Cases not falling within these results can be considered as cases with type II error.

The confidence interval for the conceptual model includes approximately 95% of the parameters estimated in 100 simulation replications (mean p-value for the drug factor in computational model: 0.14399 [CI95%: 0.1048-0.1831]; Table 1, Experiment 1). As this is the main factor considered in the clinical trial, it is an argument supporting verification of the computational model (Rodríguez-Barrios *et al*, 2008; Monleón and

Ocaña, 2006). Indeed, if the normality conditions are not fulfilled, as in this example (previous analysis using distributional test for residuals), these confidence intervals are not 95% either.

To illustrate the use of the proposed measure of stability, PVL_i, for the verification of simulation models, we performed 10 different simulation experiments using 100 replications each time (see appendix B for the methodological details) to establish the variation ranges of PVL_i using different perturbation conditions. We compared the results with p-values for drug and drug*time effects in the cases of supposing either compliance or non-compliance with the verification conditions. Table 1 summarizes our results showing the number of experimental observations, different conditions of variation versus the original simulation model, mean *PVL*, *PVL* standard deviation (Std), *PVL* range (*ul*-*ll*), % p-value<0.05 (mean p-value and 95%CI) for drug and time*drug effect and, in the last column, an observational evaluation (author criteria) of whether the simulation model is considered verified. PVL_{ul} < 20% was established as the maximum distance between the computational and conceptual models for verification.

After different simulation experiments, we observed that the values of PVL_i using *LLK*, *AIC*, *AICC* and *BIC* are very similar (Table 1, Experiment 1) so for brevity only the *LLK* results are included in Table 1, experiments 2 to 10.

The first Experiment in Table 1 is related to the verified computational model discussed previously (5) and experiments 2 to 8 are related to the variation in the global simulation error (σ). PVL_i range (*ul*-*ll*) increases a lot when the global error differs appreciably from its original value, as in Experiments 2 (σ= 15, *PVL* range: 143.950-153.099%) and 3 (σ= 0.1, *PVL* range: 82.4745-94.4352%). Consequently, non-verified simulation models are under consideration in these cases. This observation is supported by the results of %p-value<0.05 in Experiment 2 (%p-value drug<0.05 = 15% and p-value drug*time<0.05= 7%). Figure 4 represents the simulation model under the conditions of Experiment 2. The model response does not seem to be the same as in Figure 2 for the conceptual model.

Experiments 4, 5, 6 and 7 are related to the influence of increasing global errors (σ) in order to establish when PVL_{ul} reaches the 20% limit. In Experiment 4, simulations are performed using σ = 0.8663, the upper limit of the estimated global error. The result shows that PVL_{ul} < 10% in this case. However, when σ = 1.1 (Experiment 7) *PVL* range

= 10.4815-20.3600%, and when $\sigma > 1.1$ (Experiments 5 and 6; $\sigma = 2$ and $\sigma = 1.5$) $PVL_{ul} >$ 20%. From this it is clear that PVL_{ul} reaches the observed 20% limit near the value of $\sigma=1$.

The same type of experiments are possible to study the variations in a model when fixing the random parameters (β and u_i in the previous example) or other scenarios like extrapolating the model in time, studied in Experiments 8, 9 and 10. In Experiments 9 and 10, the same conditions as in the original simulation model were used but the duration of the trial was changed to 24 months and the schedule of visits to every 4 months. In Experiment 9, *LLK* was normalized with the number of total observations (n_a=349) referred to the original computational model size (n_b=399) due to the increase or decrease in likelihood. In Experiment 8 *PVL* range = (2.88799 - 14.12544)% and (3.40816 - 16.67919)% in every case all the computational models show the same behaviour as the conceptual model. Finally, the experiment was extrapolated to 48 months and the *PVL* range was (12.4083 - 24.8824)% reflecting the instability of the model, confirmed by the 85% drug p-value < 0.05 in the experiment. This indicates the tendency of the model to produce significant differences between drugs in 48-month models. So, by means of the *PVL* range and p-value we can obtain the time period for the verified model.

The *PVL* range of 0-20% is a good measure of model verification as also observed in further simulation experiments using mixed models. These simulation experiments and their verification within the framework of clinical trials are presented in previous studies (Monleón, 2005), using other data sets and different models, such as linear and non-linear mixed models. In all cases similar results were observed during the verification experiments. These results (Table 2) show a *PVL* ranges of 4-18%, 0.5-18% and 0.78-12% for three different clinical trial experiments and confirm that the range of 0 to 20 % is correct for the computational model verifying the simulation model. The value of *PVLIR* range (5) between 0 and 5 % obtained, suggests also that the computational model verifies the conceptual model, though this index has not been calculated for all the experiments.

5. Discussion

The main reason to perform simulations is to investigate new scenarios: varying some parameters and then looking how consistent the obtained results are or design new

studies where we need to know how much the results and the precision of the results vary under different designs. In this way it can be studied how robust the results are.

Unfortunately the amount of process to be performed in a simulation are huge (conceptual model, computational model, simulation replications, etc) and derives a complexity analysis of the results, especially if there is more than one parameter of interest and the detection of errors, especially during the model verification.

Our results coming from several simulation experiments of a clinical trial show that the *PVL* statistic range can be a good measure of stability to establish when a computational model verifies the underlying conceptual model.

Our results show that the *PVL* statistic range can be a good measure of stability for establishing when the computational model is verified, ensuring that the simulation computer software, programming and implementation of the conceptual model are correct. They suggest that a *PVL* range of 0-20% indicates that the computational model is verified, as an acceptable guideline. One could apply the *PVL* approach to every conceptual model that was tried during the model building process. However the simulation of the complex models are still much limited and can be cautious in the generalization on this methodology to any statistical model and simulation.

We can also verify the average simulation predictions using replications. The idea could be improved by calculating simultaneous confidence intervals for all the parameters or for all the paths, and not parameter to parameter and period to period as is frequently the case. This problem is not present in the *PVL* range stability measure, since only one value is associated with all the simulation replications and consequently improves its. This measure, presented for verification, is useful only when the conceptual model remains totally specified by a mathematical expression (as, for example, in a mixed model) which allows us to define likelihood confidence intervals. On some occasions (frequently coinciding with the most interesting cases to simulate) the conceptual model and its implementation in the shape of the conceptual model includes aspects not contemplated in these expressions (such as patient withdrawal, time between visits, fortuitous facts, queuing model, etc.). This implies that no direct confidence intervals or other statistical properties are available. In these cases, alternative approaches should be used, such as the graphical analysis summarized in Sargent (2003), as well as the opinion of the experts in the specific subject area.

Acknowledgements

The authors would like to thank professor Ludwig Fahrmeir from the Institut für Statistik at the Ludwig Maximilians University in Munich (Germany) for checking this manuscript and his wise comments. Part of this work was completed at the Institut für Statistik and we are most grateful for their hospitality.

This research was supported by a scholarship from the Department of Statistics at the University of Barcelona, grant no.: ACES-UB 2006.

Appendix A.

Mixed models theory

The general form of a mixed model is,

$$Y = X\beta + Zu + e, \qquad (A1)$$

where Y is an $nx1$ vector of observed random variables (data). X and Z are known design matrices, β is a $px1$ vector of fixed effects, u is an $mx1$ vector of random variables (random factors) and e is an $nx1$ vector of random error terms. Vectors u and e are distributed randomly with multivariate Gaussian distribution with $\mu = 0$ and covariance matrices G ($m \times m$) and R ($n \times n$) respectively. It is supposed that $\mathrm{Cov}(u,e)=0$, so $V(Y) = ZGZ' + R = V$.

If G and R *are known*, V is also known, and solutions obtained using a generalized least squares method to estimate vector β represent the *Best Linear Unbiased Estimator* (BLUE) value of the fixed parameters of the model and the solution $\hat{u} = GZ'V^{-1}(y'X\hat{\beta})$ is the value of the *Best Linear Unbiased Predictor* (BLUP) of the random effects (Searle *et al*. [22]). In reality V is unknown, so components of variance in G and R are initially estimated to obtain $\hat{V}$ which replaces V in the estimate of β and the prediction of u. In the parametric supposition, variance components and covariance can be estimated using procedures based on the log likelihood function (Hayman 1960; Harville, 1977).

In the mixed model, the likelihood is calculated from a normal distribution and its formula is (Hartley and Rao, 1967):

$$L = (2\pi)^{-\frac{n}{2}} |V|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y - X\beta)' V^{-1} (y - X\beta)\right] \qquad (A2)$$

where $\mathbf{Y} \sim N(X\boldsymbol{\beta}, V)$ with $V = ZGZ' + \sigma^2 I$

Appendix B.

Determining the number of simulation replications

For R independent replications of one random variable Y (mobility in this study) the i-th observation of the replication r may be denoted as Y_{ri} for i = 1, 2, ..., n_r and r = 1,2,...,R. In the example we have R=100.

Every simulation replication r (for r=1,…,R) produces n_r vectors, usually mutually independent. The $Y_{r1},...,Y_{rn_r}$ correspond to the n_r simulated cases in a general replication, or n if the case number is constant. Also, every vector Y_{ri} has T_{ri} value*s of* Y_{rit} (t=*1,..., *T_{ri}) that form a sequence (frequently auto-correlated) of random variables corresponding to the Y values through time. If the number of time moments is considered constant, then T_{ri}=T.

For simplicity, we consider here that Y_{ri} directly corresponds to a scalar summary of the T values corresponding to the *i-th* case during simulation replication r.

The simple average of the simulation replications of Y is defined as,

$$\hat{\theta}_r = \frac{1}{n_r}\sum_{i=1}^{n_r} Y_{ri},\ r = 1,2,3,...,R \tag{B1}$$

The general sample average is:

$$\hat{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r \tag{B2}$$

The simple simulation variance is defined as:

$$\hat{\sigma}^2(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\frac{(\hat{\theta}_r - \hat{\theta})^2}{R-1} \tag{B3}$$

The confidence interval based on the Gaussian distribution of the θ parameter with a coverage of 100(1-α)% is calculated as:

$$\hat{\theta} - \frac{t_{\alpha/2,R-1}S}{\sqrt{R}} \le \theta \le \hat{\theta} + \frac{t_{\alpha/2,R-1}S}{\sqrt{R}} \tag{B4}$$

Where the standard deviation is defined as:

$$S = \sqrt{R}\hat{\sigma}(\hat{\theta}) \tag{B5}$$

The confidence semi-interval for θ is defined as:

$$g = \frac{t_{\alpha/2,R-1}S}{\sqrt{R}}, \tag{B6}$$

Accepting the error criterion specified according to confidence interval 1-α as:

$$P(\left|\hat{\theta}-\theta\right|<\varepsilon)\geq 1-\alpha \text{ and } g-\frac{t_{\alpha/2,R-1}S}{\sqrt{R}}\leq\varepsilon \quad \text{(B7)}$$

we can deduce that,

$$R\geq(\frac{t_{\alpha/2,R-1}S}{\varepsilon})^2 \quad \text{(B8)}$$

Some authors, such as Guasch *et al* (2002) suggest formulas to calculate the number of simulation replications according to a necessary precision (model acceptability) of the response variable, and propose analysing previous replications (pilot simulation). In the continuous response variables the number of replications can be calculated as:

$$R=k\left[\frac{(g/2)}{\varepsilon}\right]^2 \quad \text{(B9)}$$

where k is the number of pilot replications, g is the semi-interval of confidence for the response variable expressed as a percentage of the response variable and ε is the required precision (1%, 5%, 10%, etc). In experiment we present, if we need to know how many replications are necessary for a precision of 15.8% and over 10 pilot simulations we observe that the semi-interval of confidence is 10% of the response variable (mobility), the number of final replications can be calculated as:

$$R=10\left[\frac{(0.158/2)}{0.1}\right]^2 = 100 \text{ replications (approx.)}$$

This way of calculating the number of replications is complicated. In many scenarios it is very difficult to calculate the confidence semi-interval; it depends on the model and the calculations performed. It is also evident that the number of replications increases if the confidence interval decreases.

References

1 Balci O (1998). Validation, verification, and accreditation. Proceedings of the 1998 Winter Simulation Conference. IEEE: Piscataway, NJ, pp 41–48.

2 El-Halimi R, Ocaña J, Ruiz De Villa MC (2004). A simulation study on the robustness of parametric inference in a nonlinear mixed modelling context (available at http://www.imub.ub.es/publications/preprints/pdf/prepri367.pdf accessed 16 July 2009).

3 Geaghan J (2002). Repeated measures example 1. (Statistical Laboratory, Dept. of Experimental Statistics, Louisiana State University, EXST 7013 Spring (available at www.stat.lsu.edu/exstweb/statlab/7013/repeated2.htm accessed 16 July 2009).

4 Guasch MA, Piera J, Casanovas J, Figueras J (2002). Modelado y Simulación: Aplicación a Procesos Logísticos, de Fabricación y de Servicios. Edicions UPC: Barcelona.

5 Hand DJ, Taylor CC (1987). Multivariate Analysis of Variance and Repeated Measures. Chapman and Hall: London.

6 Hartley HO, Rao JNK (1967). Maximum likelihood estimation for the mixed analysis of variance model. Biometrika 54: 93-108.

7 Harville DA (1977). Maximum likelihood approaches to variance component estimation and to related problems. J Amer Stat Assoc 72: 320-340.

8 B.I. Hayman (1960). Maximum likelihood estimation of genetic components of variation. Biometrics 16: 369-381.

9 Holford NHG, Hale M, Ko HC, Steimer JL, Sheiner LB, Peck CC (1999). Simulation in drug development: good practices. Georgetown University: USA (available at http://cdds.ucsf.edu/research/sddgpreport.php accessed 16 July 2009).

10 Holford NHG, Kimko HC, Monteleone JPR, Peck CC (2000). Simulation of clinical trials. Annual review of pharmacology and toxicology 40: 209-234.

11 Jenrich RL, Schluchter MD (1986). Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42: 805-820.

12 Littell RC, Milliken GA, Stroup WW, Wolfinger RD (1986). SAS System for Mixed Models. Cary NC, SAS Institute Inc: USA.

13 Monleón T (2005). Clinical trial of drugs optimization by means of discrete event simulation, its modelling, validation, monitoring and the improvement its information quality, PhD thesis, University of Barcelona, (available at http://www.tesisenxarxa.net/TDX-0112106-093218 accessed 16 July 2009).

14 Monleón T, Ocaña J (2006). Use of likelihood criteria for the verification of realisic simulation models. Book of abstracts, Dipartamento di Statistica, Probabilità e Statistiche Applicate (eds). Proceedings of the 17th Symposium of IASC on Computational Statistics (Roma) : 214-215.

15 Monleón T, Ocaña J, Vegas E, Fonseca P, Riera A, Montero J, Abbas I, J Casanovas, Cobo E, Arnaiz JA, Carne X, Gatell JM (2004). Flexible discrete events simulation of clinical trial using LeanSim. Proceedings of the 17th Symposium of IASC on Computational Statistics (Prague): 1519-1526.

16 Nance RE, Sargent RG (2002). Perspectives on the evolution of simulation. Operations Research 50: 161-172.

17 Roach PJ (1998). Verification and Validation in Computational Science and Engineering. Hermosa Publishers: Albuquerque, NM.

18 Rodríguez-Barrios JM, Serrano D, Monleón T, Caro C (2008). Discrete-event simulation models in the economic evaluation of health technologies and health products. Gaceta Sanitaria 22(2): 151-161.

19 Sargent RG (2003). Verification and validation of simulation models. Chick S, Sanchez PJ, Ferrin D, Morrice DJ (eds). Proceedings of the 2003 Winter Simulation Conference IEEE: Piscataway, NJ, pp 37-48.

20 Schlesinger S (1979). Terminology for model credibility. Simulation 32(3):103-104.

21 Searle SR, Casella G, Mc Culloch CH (1992). Variance components. Wiley: New York.

22 Thacker BH, Anderson MC, Senseny PE, Rodriguez EA (2005). The role of nondeterminism in model verification and validation. International Journal of Materials and Product Technology 25(1-2): 144-163.

23 Whiner RG, Balci O (1989). Guideline for selecting and using simulation model verification techniques. Proceedings of the 1989 Winter Simulation Conference IEEE: Piscataway, NJ, pp 559-568.

Figures and Tables

Figure 1

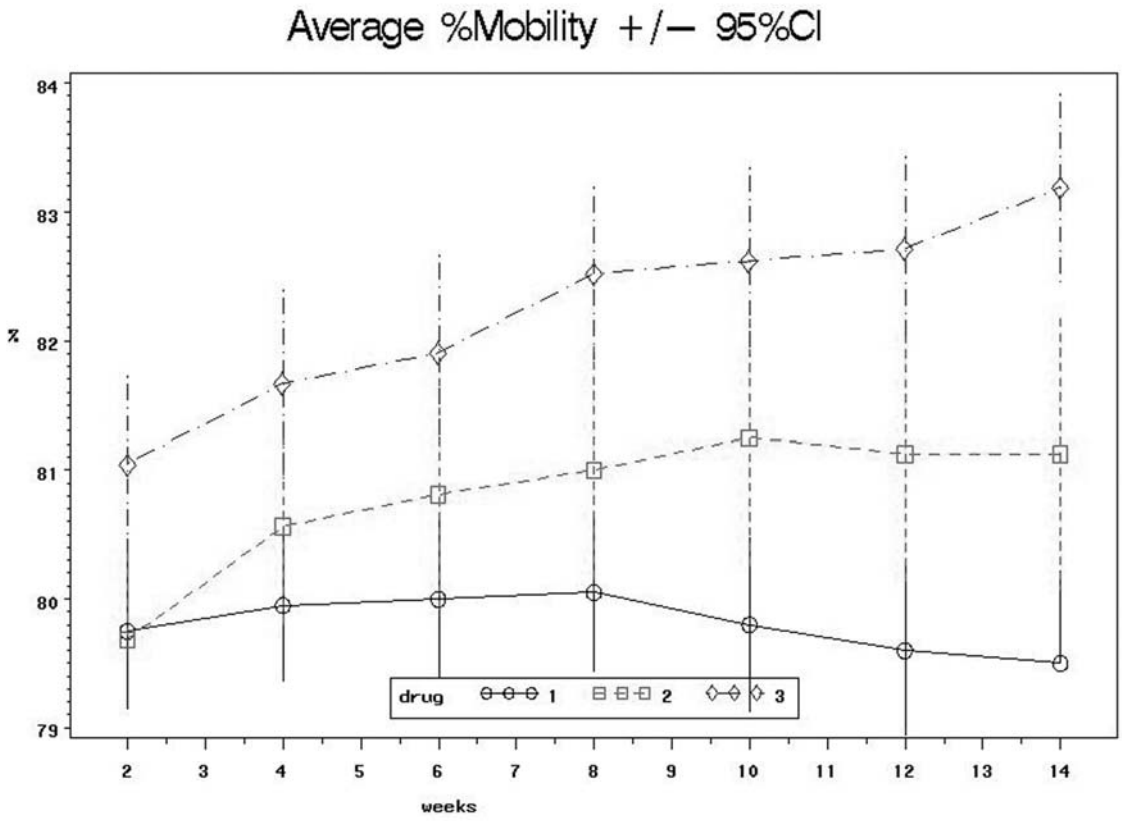

Figure 2

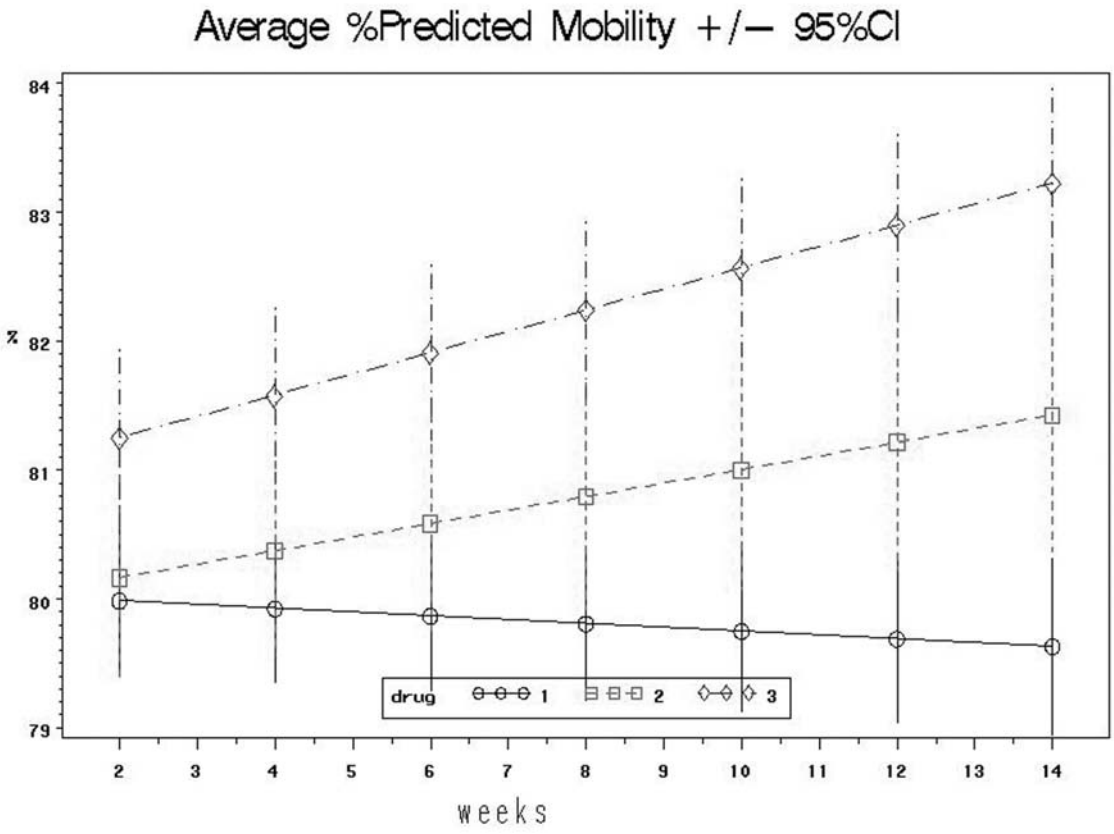
Average %Predicted Mobility +/- 95%CI
84
83
82
81
80
79
%
drug
1
2
3
2
3
4
5
6
7
8
9
10
11
12
13
14
weeks

Figure 3

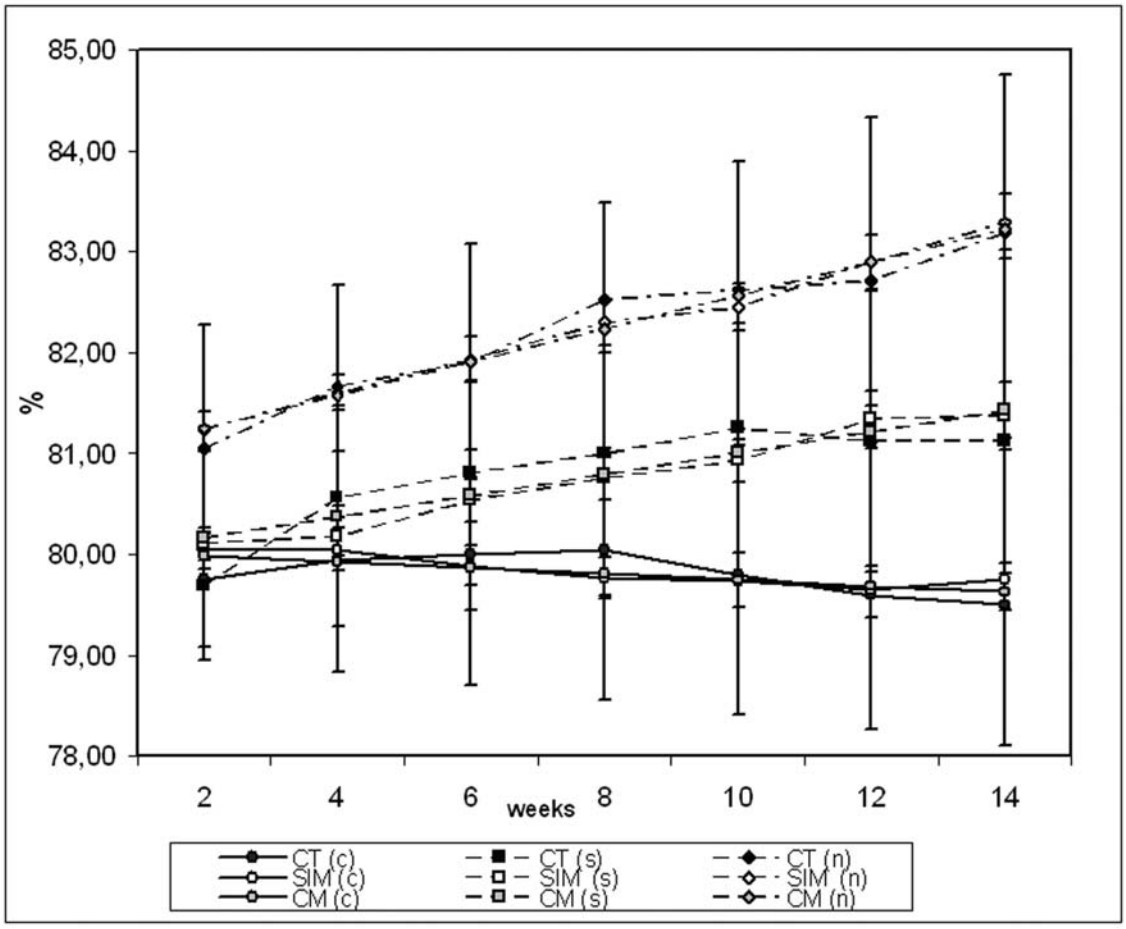
85,00
84,00
83,00
82,00
81,00
80,00
79,00
78,00
%
2
4
6
8
10
12
14
weeks
CT (c)
CT (s)
CT (n)
SIM (c)
SIM (s)
SIM (n)
CM (c)
CM (s)
CM (n)

Figure 4

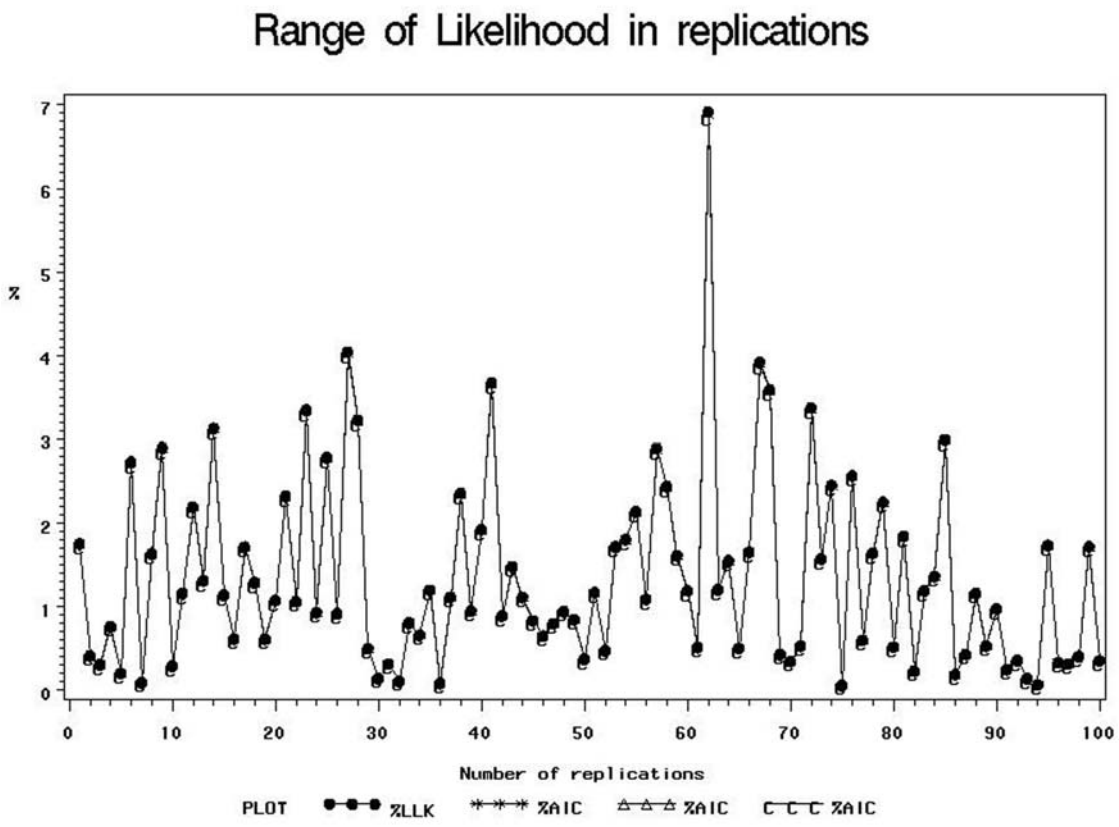
Range of Likelihood in replications
%
0
1
2
3
4
5
6
7
0
10
20
30
40
50
60
70
80
90
100
Number of replications
PLOT
%LLK
%AIC
%AIC
%AIC

Figure 5

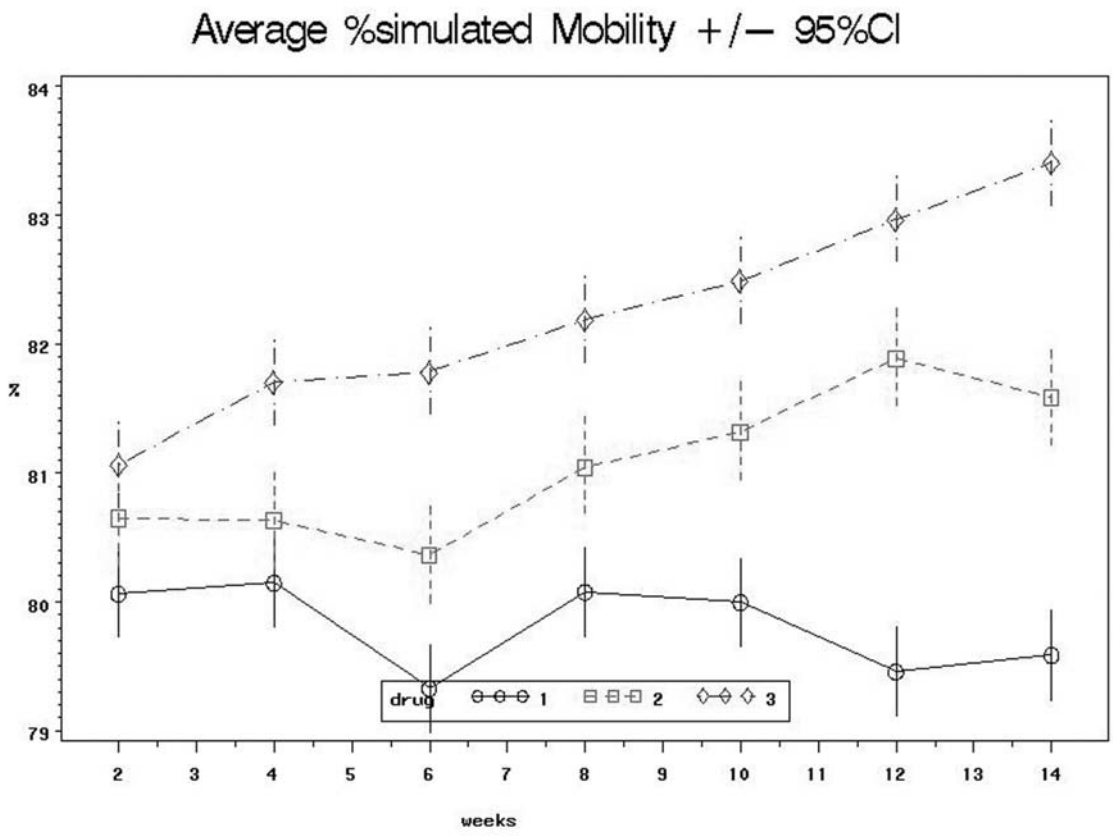
Average %simulated Mobility +/- 95%CI
84
83
82
81
80
79
%
2
3
4
5
6
7
8
9
10
11
12
13
14
weeks
drug
1
2
3

Figure 6

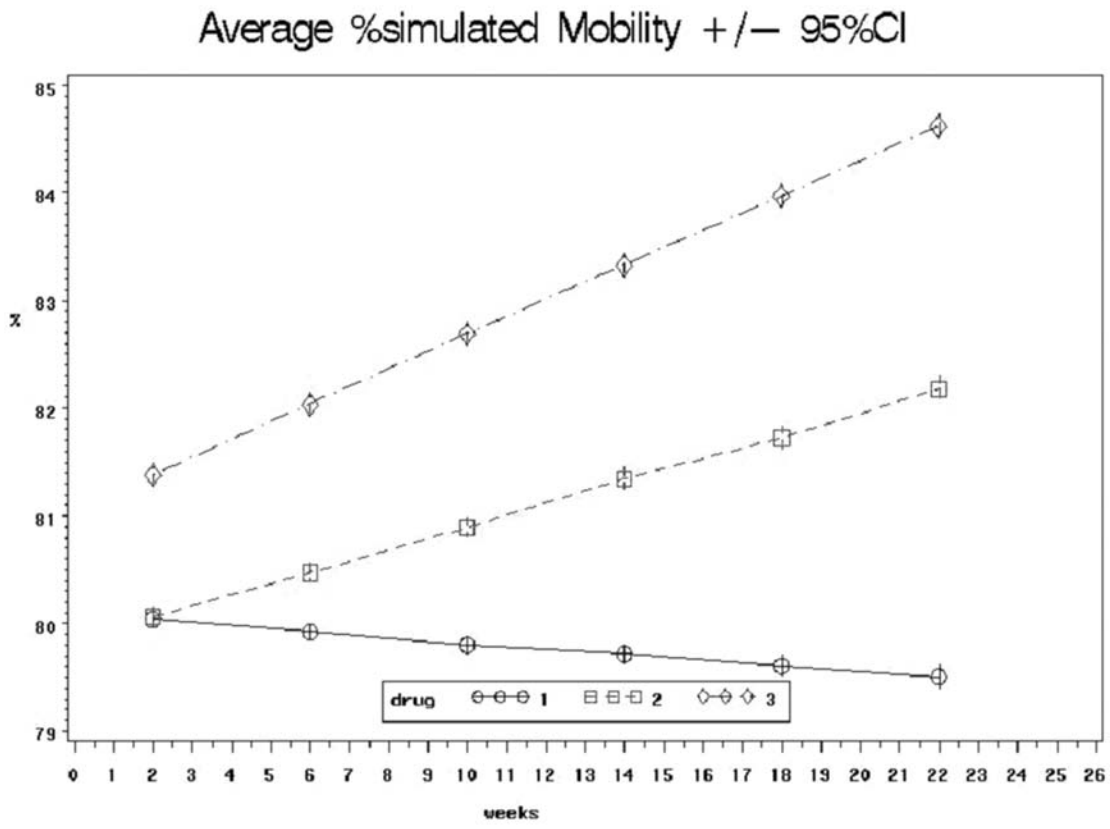
Average %simulated Mobility +/− 95%CI
%
85
84
83
82
81
80
79
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
weeks
drug 1 2 3

Figure 7

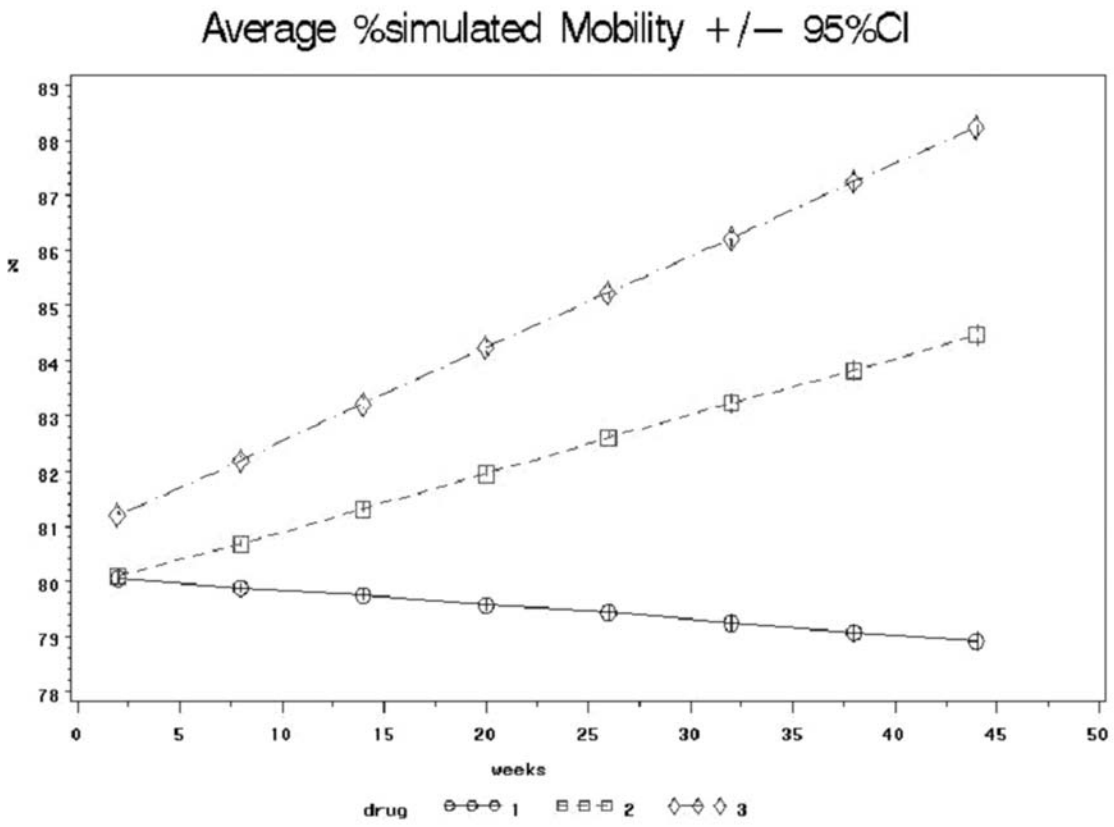
Average %simulated Mobility +/— 95%CI
%
89
88
87
86
85
84
83
82
81
80
79
78
0
5
10
15
20
25
30
35
40
45
50
weeks
drug
1
2
3

Table 1

n_e	n_b	Condition*	*PVL* Mean	*PVL* Stv	*PVL* Range (lower limit (*ll*) - upper limit (u*l*))	% p-value drug<0.05 Mean p-value drug [95%CI p-value drug]	% p-value time x d rug<0.05 Mean p-value time x drug [95%CI p-value time x drug]	Verified model using author criteria
1	399	Simulation model σ= 0.79	*LLK*:2.0334 *AIC*: 2.0213 *AICC*: 2.0211 *BIC*: 2.0991	1.3763 1.3680 1.3680 1.3591	0.0265-6.6730 0.0205-6.6329 0.0203-6.6327 0.0224-6.5902	47 0.14399 [0.1048-0.1831]	82 0.0356 [0.0198-0.0514]	YES
2	399	σ= 15	*LLK*:149.1466	2.05774	143.950-153.099	15 nd nd	7 0.3809420 [0.3235631 - 0.4383209]	NO
3	399	σ= 0.1	*LLK*: 89.57708	2.35537	82.4745-94.4352	58 nd nd	86 0.1076265 [0.0744647 - 0.1407884]	NO
4	399	σ= 0.8663 (CI upper limit)	*LLK*: 3.809628	2.06651	0.0470112- 8.9070551	47 0.1824055 [0.1330486 - 0.2317625]	82 nd	YES

5	399	σ= 2	*LLK*: 42.63919	2.12804	35.5333 - 47.7546	59 0.1029978 [0.0742355 - 0.1317601]	65 nd	NO
6	399	σ= 1.5	*LLK*: 28.88558	1.82833	24.3049- 32.6432	55 0.1080756 [0.0726793 - 0.1434720]	68 nd	NO
7	399	σ= 1.1	*LLK*: 14.67672	2.02814	10.4815-20.3600	54 0.1344879 [0.0954679 - 0.1735080]	83 0.0383866 [0.0187369 - 0.0580363]	YES
8	349	Simulation model σ= 0.79 Time = 24 Visits: 4 months	*LLK*: 7.45420	1.90049	2.88799 - 14.12544	62 0.1088210 [0.0706885 - 0.1469536]	84 0.0269640 [0.0154371- 0.0384909]	YES
9	349	Simulation model σ= 0.79 Time = 24 Visits: 4 months	*LLK***: 8.800693	2.24441	3.40816- 16.67919	62 0.1088210 [0.0706885 - 0.1469536]	84 0.0269640 [0.0154371- 0.0384909]	YES
10	456	Simulation model σ= 0.79 Time = 46 Visits: 6 months	*LLK***: 18.50607	2.45380	12.4083 - 24.8824	85 0.0185545 [0.0118431 - 0.0252659]	95 0.0118831 [0.0064708 -0.0172954]	NO

Table 2

Trials	Clinical trial Phase	Indication	Replications (*n*)	Trial type[1]	Mixed model type [2]	*PVL* range (*ll*-*ul*) for the estimated variability model
1	III	ARTHRITIS	100	R	L	0.03 to 6.7 %
2	I	ASTHMA	40	R	NL	4 - 18%
3	IV	AIDS	30	R	L	0.5 - 18%
4	II	Non real	40 (4 scenarios)	P	NL	0.78 - 12%

Captions for Figures and Tables

Figure 1: Percentage of mobility for each treatment (1: control; 2: standard; 3: new) in the arthritis clinical trial (mean +/- 95% CI) using real data.

Figure 2: Predicted percentage of mobility for each treatment (1: control; 2: standard; 3: new) in the arthritis clinical trial (mean +/- 95% CI) using the linear mixed model (conceptual model).

Figure 3: Results of the average paths of % mobility during the clinical trial for the 3 treatments under study (c: control (1); s: standard (2); n: new (3)). These paths correspond to the conceptual and simulation models. The confidence intervals are also represented (95% CI) for each of them. CT: clinical trial data; SIM: simulation data; CM: conceptual mixed model data.

Figure 4: Percentage of *PVL* between 100 simulation replications and the conceptual model using the computational model ($\sigma = 0.7953$) (PVL_{ul}: 6.65 for *LLK*, 6.61 for *AIC*, 6.61 for *AICC*, 6.57 for *BIC*).

Figure 5: Results of the average paths of % mobility during the clinical trial for the 3 treatments under study (1: control; 2: standard; 3: new). These paths correspond to the simulation model when σ(model)= 15. The confidence intervals are also represented (95% CI) for each of them.

Figure 6: Results of the average paths of % mobility during the clinical trial for the 3 treatments under study (1: control; 2: standard; 3: new). These paths correspond to the simulation model. The confidence intervals are also represented (95% CI) for each of them. CT: clinical trial data;

SIM: simulation data; CM: conceptual mixed model data. Extrapolation of the original conceptual model with the same number of visits and time = 24 months, every 4 months.

Figure 7: Results of the average paths of % mobility during the clinical trial for the 3 treatments under study (1: control; 2: standard; 2: new). These paths correspond to the simulation model. The confidence intervals are also represented (95% CI) for each of them. CT: clinical trial data; SIM: simulation data; CM: conceptual mixed model data. Extrapolation of the conceptual model with the same number of visits and time = 48 months, every 6 months.

Table 1: Results of the different simulations with the mobility clinical trial, using different conditions (* Variation with respect to conceptual mode, ** The number of observations in the conceptual model and in the computational model were different) n_e: Number of experiment; n_b: num. observations per experiment (case x time); nd: not determined
(1) P: prospective trial; R: retrospective trial
(2) L: Linear mixed model; NL Non-linear mixed model

Table 2: Summary of the range of variation obtained in the test of verification based on comparing the *range* of *PVL* values between simulation replications and the conceptual model in the experiments in Monleón (2005)