The design of simulation studies in medical statistics

Andrea Burton[1,2,*,†], Douglas G. Altman[1], Patrick Royston[1,3] and Roger L. Holder[4]

[1]*Cancer Research UK/NHS Centre for Statistics in Medicine, Oxford, U.K.*
[2]*Cancer Research UK Clinical Trials Unit, University of Birmingham, Birmingham, U.K.*
[3]*MRC Clinical Trials Unit, London, U.K.*
[4]*Department of Primary Care and General Practice, University of Birmingham, Birmingham, U.K.*

SUMMARY

Simulation studies use computer intensive procedures to assess the performance of a variety of statistical methods in relation to a known truth. Such evaluation cannot be achieved with studies of real data alone. Designing high-quality simulations that reflect the complex situations seen in practice, such as in prognostic factors studies, is not a simple process. Unfortunately, very few published simulation studies provide sufficient details to allow readers to understand fully all the processes required to design a simulation study. When planning a simulation study, it is recommended that a detailed protocol be produced, giving full details of how the study will be performed, analysed and reported. This paper details the important considerations necessary when designing any simulation study, including defining specific objectives of the study, determining the procedures for generating the data sets and the number of simulations to perform. A checklist highlighting the important considerations when designing a simulation study is provided. A small review of the literature identifies the current practices within published simulation studies. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: simulation study; design; protocol; bias; mean square error; coverage

1. INTRODUCTION

Simulation studies use computer intensive procedures to test particular hypotheses and assess the appropriateness and accuracy of a variety of statistical methods in relation to the known truth. These techniques provide empirical estimation of the sampling distribution of the parameters of interest that could not be achieved from a single study and enable the estimation of accuracy measures, such as the bias in the estimates of interest, as the truth is known [1]. Simulation studies are increasingly being used in the medical literature for a wide variety of situations, (e.g. References

*Correspondence to: Andrea Burton, Cancer Research UK/NHS Centre for Statistics in Medicine, Wolfson College Annexe, Linton Road, Oxford OX2 6UD, U.K.
†E-mail: andrea.burton@cancer.org.uk

[2–4]). In addition, simulations can be used as instructional tools to help with the understanding of many statistical concepts [5, 6].

Designing high-quality simulations that reflect the complex situations seen in practice, such as in randomized controlled trials or prognostic factor studies, is not a simple process. Simulation studies should be designed with similar rigour to any real data study, since the results are expected to represent the results of simultaneously performing many real studies. Unfortunately, in very few published simulation studies are sufficient details provided to assess the integrity of the study design or allow readers to understand fully all the processes required when designing their own simulation study. Performing any simulation study should involve careful consideration of all design aspects of the study prior to commencement of the study from establishing the aims of the study, the procedures for performing and analysing the simulation study through to the presentation of any results obtained. These are generic issues that should be considered irrespective of the type of simulation study but there may also be further criteria specific to the area of interest, for example survival data.

It is important for researchers to know the criteria for designing a good quality simulation study. The aim of this paper is to provide a comprehensive evaluation of the generic issues to consider when performing any simulation study, together with a simple checklist for researchers to follow to help improve the design, conduct and reporting of future simulation studies. The basic concepts underpinning the important considerations will be discussed, but full technical details are not provided and the readers are directed towards the literature (e.g. References [7, 8]). General considerations are addressed rather than the specific considerations for particular situations where simulations are extremely useful, such as in Bayesian clinical trials designs (e.g. Reference [9]), sample size determination (e.g. References [3, 10]), or in studies of missing data (e.g. Reference [4]). A small formal review of the current practice within published simulation studies is also presented.

2. ISSUES TO CONSIDER WHEN DESIGNING A SIMULATION STUDY

When planning any simulation study, as with randomized controlled trials, a detailed protocol should be produced giving full details of how the study is to be performed, analysed and reported. The protocol should document the specific objectives for the simulation study and the procedures for generating multivariate data sets and, if relevant, with censored survival times. The choices for the different scenarios to be considered, for example different sample sizes, and the methods that will be evaluated should also be included in the protocol together with the number of simulations that will be performed. It is also important to give careful consideration to which data and results will be stored from each run, and which summary measures of performance will be used. If an aim of the study is to judge which is the best of two or more methods, then the criteria should be pre-specified in the protocol, where possible. The rationale behind all the decisions made throughout the design stage should be included in the protocol.

Each of the preceding considerations will be discussed in more detail in the following sections. A checklist of the important issues that require consideration when designing a simulation study is provided in Figure 1.

2.1. Clearly defined aims and objectives

Establishing clearly defined aims for the simulation study prior to its commencement is an essential part of any research. This focuses the study and avoids unnecessary repetition and time wasting from having to repeat simulations when new aims are conceptualized.
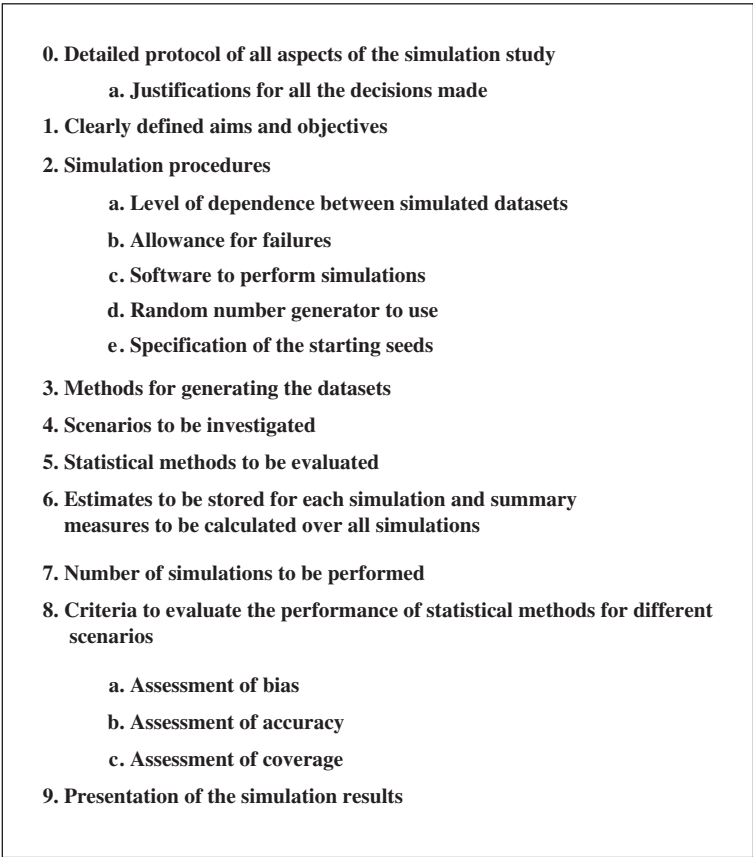
0. Detailed protocol of all aspects of the simulation study

    a. Justifications for all the decisions made

1. Clearly defined aims and objectives

2. Simulation procedures

    a. Level of dependence between simulated datasets

    b. Allowance for failures

    c. Software to perform simulations

    d. Random number generator to use

    e. Specification of the starting seeds

3. Methods for generating the datasets

4. Scenarios to be investigated

5. Statistical methods to be evaluated

6. Estimates to be stored for each simulation and summary measures to be calculated over all simulations

7. Number of simulations to be performed

8. Criteria to evaluate the performance of statistical methods for different scenarios

    a. Assessment of bias

    b. Assessment of accuracy

    c. Assessment of coverage

9. Presentation of the simulation results

Figure 1. Important considerations when designing any simulation study.

2.2. Simulation procedures

Once the aims and objectives have been formalized, the procedures for performing the simulations can be considered including the level of dependence between simulations, the allowance for failures, the choice of random number generator, starting seeds and the software package to be used. The statistical software package must be able to handle the complexities involved in the proposed simulation study and have a reliable random number generator.

All simulation studies involve generating several independent simulated data sets. These generated data sets should also be completely independent for the different scenarios considered, such as different sample sizes. However, when more than one statistical methodology is being investigated, there is an added complication of determining the level of dependence of the simulated data sets for the different methods, although still retaining independent data sets for each scenario studied. Two feasible simulation strategies are possible. Firstly, fully independent simulated data sets involve generating a completely different set of independent data sets for each method and scenario considered. Secondly, moderately independent simulations use the same set of simulated

independent data sets to compare a variety of statistical methods for the same scenario, but a different set of data sets is generated for each scenario investigated. These moderately dependent samples are like a matched pair design where the within sample variability is eliminated and therefore are sensitive to detecting any differences between methods. The relationship between the generated samples should form an important consideration when designing the study.

The simulation procedures should have some allowance for failing to estimate the outcome or parameter of interest, e.g. due to rare events or lack of convergence of models, to avoid premature stopping of the study. The simulations can be set up so that a failed sample is discarded and the whole process is repeated. The number of failures that occur should be recorded to gauge how likely this could happen in practice in order to judge whether the applied statistical procedure can reliably be used in the situation being investigated. If many failures occur for a particular scenario causing the early termination of the simulation study, researchers must consider whether in their situation the failures would lead to bias, and hence unacceptable results, or unbiased but imprecise results in order to determine the usefulness of the results from the partial set of completed simulations. Failures for some simulations may result in a *post hoc* change of the protocol to omit scenarios, which cannot be simulated reliably.

2.2.1. Random number generation. A fundamental part of any simulation study is the ability to generate random numbers. The many different types of random number generator have been detailed elsewhere [11, 12]. Any random number generator should be long in sequence before repetition and subsets of the random number sequence should be independent of each other [13]. A variety of statistical tests for randomness exist, including Marsaglia's Diehard battery of tests for randomness [14], which each random number generator must pass before it can be reliably adopted as a means of generating random numbers.

A random number generator must be able to reproduce the identical set of random numbers when the same starting value, known as a seed, is specified [13]. This is also essential when performing simulation studies to enable the generated data sets and hence results to be reproduced, if necessary, for monitoring purposes. The specification of the starting seed also facilitates the choice of simulation strategy. The simulations will be fully independent if completely different starting seeds are used to generate the data sets for each scenario and method combination considered or moderately independent if the same starting seeds are used to compare various methods for the same scenario but different seeds are employed for alternative scenarios. Any simulation strategy involves running several independent simulations for the same scenario, known as parallel simulations, which require independent sequences of random numbers. Random numbers can be generated for parallel simulations by setting different starting values for the individual simulations that are greater than the number of random numbers required for each simulation, which reduces the possibility of correlations between samples [13]. For example, if each simulated data set had a sample size of 500, then each of the 250, say, simulations would require 500 random numbers, therefore the starting seed for each simulation should be separated by at least 500.

2.3. Methods for generating the data sets

The methods for obtaining simulated data sets should be carefully considered and a thorough description provided in both the protocol and any subsequent articles published. Simulating data sets requires an assumed distribution for the data and full specification of the required parameters.

The simulated data sets should have some resemblance to reality for the results to be generalizable to real situations and have any credibility. A good approach is to use a real data set as the motivating example and hence the data can be simulated to closely represent the structure of this real data set. The actual observed covariate data could be used and only the outcome data generated or just certain aspects, such as the covariate correlation structure, could be borrowed. Alternatively, the specifications could be arbitrary, but the generated data set may be criticized for not resembling realistic situations. The rationale for any choices made regarding the distributions of the data, parameters of any statistical models and the covariate correlation structure used to generate the data set should accompany their specifications. The generated data should be verified to ensure they resemble what is being simulated, for example using summary measures for the covariate distributions, Kaplan–Meier survival curves for survival data or fitting appropriate regression models.

2.3.1. Univariate data. Simple situations may involve generating a vector of random numbers sampled from a known distribution. Demirtas [15] provides procedures for obtaining a variety of univariate distributions from initial values generated from the uniform distribution, if the required distribution is unavailable within the statistical package.

2.3.2. Multivariate data. Generating multivariate data involves the additional specification of correlations between covariates unless the covariates are assumed fully independent, which is unlikely in practice. The specification of the means and associated covariance matrix is more straightforward if based on real data, especially with a large number of covariates, and the generated data will reflect reality. Conversely, the choice of the correlations between covariates can be arbitrary but it is often problematic to determine what are valid relationships. The simplest approach to generate multivariate covariate data with a specified mean and correlation structure is to assume a multivariate normal distribution. Any continuous but non-normally distributed variables in the real data should be transformed to make the assumption of normality more appropriate. Binary variables can be generated as latent normal, i.e. generated as continuous variables and then dichotomized, but the covariate correlation structure used to generate the continuous variable needs to be adjusted to provide the correct correlation with the resulting binary variable [16]. For example, the correction factor for a continuous variable that is to be dichotomized with a 50:50 split is 0.80, suggesting that the correlation between a continuous variable and a binary variable is 20 per cent less than the correlation between two continuous variables [16].

2.3.3. Time to event data. When the outcome is time to an event, such as in prognostic modelling, several additional considerations must be addressed. The simulations require the specification of a model for the multivariate covariate data and a distribution for the survival data, which may be censored. In order to simulate censored survival data, two survival distributions are required, one for the uncensored survival times that would be observed if the follow-up had been sufficiently long to reach the event and another representing the censoring mechanism.

The empirical survival distribution from a similar real data set would provide a reasonable choice for the survival distribution. The uncensored survival distribution could be generated to depend on a set of covariates with a specified relationship with survival, which represents the true prognostic importance of each covariate. Time-dependent covariates could also be simulated and incorporated following the procedures described by Mackenzie and Abrahamowicz [17]. Bender *et al.* [18] discuss the generation of survival times from a variety of survival distributions including

the exponential for constant hazards, Weibull for monotone increasing or decreasing hazards and Gompertz for modelling human mortality, in particular for use with the Cox proportional hazards model.

Random non-informative right censoring with a specified proportion of censored observations can be generated in a similar manner to the uncensored survival times by assuming a particular distribution for the censoring times, such as an exponential, Weibull or uniform distribution but without including any covariates. Determining the parameters of the censoring distribution given the censoring probability is often achieved by iteration. However, Halabi and Singh [10] provide formulas for achieving this for standard survival and censoring distributions. The censoring mechanism can also be extended to incorporate dependent, informative censoring [19].

The survival times incorporating both events and censored observations are calculated for each case by combining the uncensored survival times and the censoring times. If the uncensored survival time for a case is less than or equal to the censored time, then the event is considered to be observed and the survival time equals the uncensored survival time, otherwise the event is considered censored and the survival time equals the censored time.

2.4. Scenarios to be investigated and methods for evaluation

Simulation studies usually examine the properties of one or more statistical methods in several scenarios defined by values of various factors such as sample size and proportion of censoring. These factors are generally examined in a fully factorial arrangement. The number of scenarios to be investigated and the methods for evaluation must be determined and justifications for these choices provided in the protocol. The scenarios investigated should aim to reflect the most common circumstances and if possible cover the range of plausible parameter values. The number of scenarios and statistical methods to investigate will depend on the study objectives but may be constrained by the amount of time available, the efficiency of the programming language and the speed and availability of several computers to run simulations simultaneously [20].

2.5. Estimates obtained from each simulation

It is essential to plan how the estimates will be stored after each simulation. Storing estimates enables consistency checks to be performed and allows for the identification of any errors or outlying values and the exploration of any trends and patterns within the individual simulations that may not be observed from the summary measure alone. Storing estimates also allows different ways of summarizing the estimates to be calculated retrospectively, if necessary, without the need to repeat all the simulations. A thorough consideration at the design stage of the possible estimates that may be of interest can ensure that all the required estimates are included, analysed and the results stored, and will avoid the risk of needing to repeat simulations. The estimate of interest, $\hat{\beta}_i$, could include the mean value of a variable, the parameter estimate after fitting a regression model, the log hazard ratio for survival models or the log odds ratios for logistic regression models. An associated within simulation standard error (SE) for the estimate of interest, $\mathrm{SE}(\hat{\beta}_i)$, is generally required.

It is also important to establish how to summarize these estimates once all simulations have been performed. Many published simulation studies report the average estimate of interest over the B simulations performed, e.g. $\bar{\hat{\beta}} = \sum_{i=1}^{B} \hat{\beta}_i / B$ as a measure of the true estimate of interest. Simulations are generally designed to mimic the results that could have been obtained from a

single study and therefore an assessment of the uncertainty in the estimate of interest between simulations, denoted $\text{SE}(\hat{\beta})$, is usually the empirical SE, calculated as the standard deviation of the estimates of interest from all simulations, $\sqrt{[1/(B-1)]\sum_{i=1}^{B}(\hat{\beta}_i - \bar{\hat{\beta}})^2}$. Alternatively, the average of the estimated within simulation SE for the estimate of interest $\sum_{i=1}^{B}\text{SE}(\hat{\beta}_i)/B$ could be used. The empirical SE should be close to the average of the estimated within simulation SE if the estimates are unbiased [21] and therefore, it may be sensible to consider both estimates of uncertainty. Alternatively, if using the mean and SE of the estimates over all simulations is not considered appropriate then non-parametric summary measures using quantiles of the distribution could be obtained.

2.6. Number of simulations required

The number of simulations to perform can be based on the accuracy of an estimate of interest, e.g. a regression coefficient, as with determining the sample size for any study [22, 23]. The number of simulations (B) can be calculated using the following equation:

$$B = \left(\frac{Z_{1-(\alpha/2)}\sigma}{\delta} \right)^2 \tag{1}$$

where δ is the specified level of accuracy of the estimate of interest you are willing to accept, i.e. the permissible difference from the true value β, $Z_{1-(\alpha/2)}$ is the $1-(\alpha/2)$ quantile of the standard normal distribution and σ^2 is the variance for the parameter of interest [22, 23]. A realistic estimate of the variance may be obtained from real data if the simulations are based on a real data set and are trying to maintain the same amount of variability. If the variance is unknown or cannot be estimated reliably then it may be possible to perform an identical simulation study to obtain realistic estimates for the variance or consider the measure of accuracy as a percentage of the SE. For example, if the variance from fitting a single covariate in a Cox regression model was 0.0166, then the number of simulations required to produce an estimate to within 5 per cent accuracy of the true coefficient of 0.349 with a 5 per cent significance level would be only 209. To estimate the regression coefficient to within 1 per cent of the true value would require 5236 simulations. Alternatively, the number of simulations could be determined based on the power $(1-\theta)$ to detect a specific difference from the true value as significant [22], such that

$$B = \left(\frac{(Z_{1-(\alpha/2)} + Z_{1-\theta})\sigma}{\delta} \right)^2$$

In fact, this formula is equivalent to equation (1) if the power to detect a specified difference is assumed to be 50 per cent.

The number of simulations to perform is thus dependent on the true value of the estimate of interest, the variability of the estimate of interest, and the required accuracy. For example, more simulations are needed if the regression coefficient is small or the estimate has little variability. Increasing the number of simulations will reduce the SE of the simulation process, i.e. $\text{SE}(\hat{\beta})/\sqrt{B}$, but this can be computational expensive and therefore variance reduction techniques could be employed [24]. The rationale for the number of simulations to perform should be included in the protocol.

2.7. *Evaluating the performance of statistical methods for different scenarios*

After the simulations have been performed, the required estimates stored after each replication and summary measures calculated, it is necessary to consider the criteria for evaluating the performance of the obtained results from the different scenarios or statistical approaches being studied. The comparison of the simulated results with the true values used to simulate the data provides a measure of the performance and associated precision of the simulation process. Performance measures that are often used include an assessment of bias, accuracy and coverage. Collins *et al.* [4] emphasized the importance of examining more than one performance criterion such as mean square error (MSE), coverage and width of the confidence intervals in addition to bias, as results may vary across criteria. In general, the expectation of the simulated estimates is the main interest and hence the average of the estimates over all simulations is used to calculate accuracy measures, such as the bias. When judging the performance of different methods, there is a trade-off between the amount of bias and the variability. Some argue that having less bias is more crucial than producing a valid estimate of sampling variance [25]. However, methods that result in an unbiased estimate with large variability or conversely a biased estimate with little variability may be considered of little practical use. The most commonly used performance measures are considered in turn. Table I provides a summary of the most applicable performance measures and formulas.

Table I. Performance measures for evaluating different methods.

Evaluation criteria	Formula
BIAS	
Bias	$\delta = \bar{\hat{\beta}} - \beta$
Percentage bias	$\left(\dfrac{\bar{\hat{\beta}} - \beta}{\beta} \right) * 100$
Standardized bias	$\left(\dfrac{\bar{\hat{\beta}} - \beta}{\mathrm{SE}(\hat{\beta})} \right) * 100$
ACCURACY	
Mean square error	$(\bar{\hat{\beta}} - \beta)^2 + (\mathrm{SE}(\hat{\beta}))^2$
COVERAGE	Proportion of times the $100(1-\alpha)\%$ confidence interval $\hat{\beta}_i \pm Z_{1-\alpha/2}\mathrm{SE}(\hat{\beta}_i)$ include β, for $i = 1, \ldots, B$.
Average $100(1-\alpha)\%$ confidence interval length	$\dfrac{\sum_{i=1}^{B} 2Z_{1-\alpha/2}\mathrm{SE}(\hat{\beta}_i)}{B}$

Key: β is the true value for estimate of interest, $\bar{\hat{\beta}} = \sum_{i=1}^{B} \hat{\beta}_i / B$, B is the number of simulations performed, $\hat{\beta}_i$ is the estimate of interest within each of the $i = 1, \ldots, B$ simulations, $\mathrm{SE}(\hat{\beta})$ is the empirical SE of the estimate of interest over all simulations, $\mathrm{SE}(\hat{\beta}_i)$ is the SE of the estimate of interest within each simulation and $Z_{1-(\alpha/2)}$ is the $1 - (\alpha/2)$ quantile of the standard normal distribution.

2.7.1. Assessment of bias. The bias is the deviation in an estimate from the true quantity, which can indicate the performance of the methods being assessed. One assessment of bias is the difference between the average estimate and the true value, i.e. $\delta = \bar{\hat{\beta}} - \beta$ (Table I). The amount of bias that is considered troublesome has varied from $\frac{1}{2}\mathrm{SE}(\hat{\beta})$ [21] to $2\mathrm{SE}(\hat{\beta})$ [26]. Another approach is to calculate the bias as a percentage of the true value (Table I), providing the true value does not equal to zero. The percentage bias could have a detrimental effect on the results if the bias is greater than the amount specified when determining the number of simulations required. Alternatively, the bias as a percentage of the $\mathrm{SE}(\hat{\beta})$ (Table I) can be more informative, as the consequence of the bias depends on the size of the uncertainty in the parameter estimate [4]. A standardized bias of greater than 40 per cent in either direction has been shown to have noticeable adverse impact on the efficiency, coverage and error rates [4].

Testing the significance of the amount of bias in the estimates [27] or obtaining a 95 per cent confidence interval using the average parameter estimate, $\bar{\hat{\beta}}$, seem counterintuitive, since these statistics are based on the number of simulations through the $\mathrm{SE}(\bar{\hat{\beta}}) = \mathrm{SE}(\hat{\beta})/\sqrt{B}$ and hence these statistics can be improved or penalized by changing the number of simulations performed (see Section 2.6). Collins *et al.* [4] warned that with a large number of simulations, the bias may be deemed statistically significant but not be practically significant. Therefore do not rely solely on the *p*-value but consider the amount of bias as well.

2.7.2. Assessment of accuracy. The MSE provides a useful measure of the overall accuracy (Table I), as it incorporates both measures of bias and variability [4]. The square root of the MSE transforms the MSE back onto the same scale as the parameter [4].

2.7.3. Power, type I and II errors. The empirical power of a test, where relevant, can be determined as the proportion of simulation samples in which the null hypothesis of no effect is rejected at the nominal, usually 5 per cent, significance level, when the null hypothesis is false (e.g. References [3, 28]). Hence the empirical type II error rate is 1-power. The empirical type I error can be calculated as the proportion of *p*-values from testing the null hypothesis of no difference on each simulated sample that are less than the nominal 5 per cent significance level, when the null hypothesis is true (e.g. Reference [29]).

2.7.4. Assessment of coverage. The coverage of a confidence interval is the proportion of times that the obtained confidence interval contains the true specified parameter value (Table I). The coverage should be approximately equal to the nominal coverage rate, e.g. 95 per cent of samples for 95 per cent confidence intervals, to properly control the type I error rate for testing a null hypothesis of no effect [4]. Over-coverage, where the coverage rates are above 95 per cent, suggests that the results are too conservative as more simulations will not find a significant result when there is a true effect thus leading to a loss of statistical power with too many type II errors. In contrast, under-coverage, where the coverage rates are lower than 95 per cent, is unacceptable as it indicates over-confidence in the estimates since more simulations will incorrectly detect a significant result, which leads to higher than expected type I errors. A possible criterion for acceptability of the coverage is that the coverage should not fall outside of approximately two SEs of the nominal coverage probability (p), $\mathrm{SE}(p) = \sqrt{p(1-p)/B}$ [27]. For example, if 95 per cent confidence intervals are calculated using 1000 independent simulations then $\mathrm{SE}(\hat{p})$ is

0.006892 and hence between 936 and 964 of the confidence intervals should include the true value.

The average length of the 95 per cent confidence interval for the parameter estimate $\hat{\beta}$ (Table I) is often considered as an evaluation tool in simulation studies (e.g. References [4, 30]). If the parameter estimates are relatively unbiased then narrower confidence intervals imply more precise estimates, suggesting gains in efficiency and power [30].

2.8. Presentation of the simulation results

Simulation studies can generate a substantial amount of results that need to be summarized and displayed in a clear and concise manner for the conclusions to be understood. The appropriate format is highly dependent on the nature of the information presented and hence there is a lack of a consistency in the literature. Structuring a report of any simulation study using separate subheadings for the objectives, methods, results and discussion provides clarity and can aid interpretation.

3. REVIEW OF CURRENT PRACTICE

A small formal review of articles published during 2004 in the Statistics in Medicine journal that included 'simulation' in the title, abstract or as a keyword was carried out to identify the current practices within published simulation studies. Of all 270 articles published in 2004, 58 (21 per cent) were identified as reporting a simulation study; their characteristics are summarized in Table II.

The specifics of the random number generator and the choice of starting seeds were generally omitted from the publications. Only one of the 58 articles explicitly stated the random number generator that was used; drand48 on the Unix/LINUX system [31]. Twenty-two articles gave some indication of the software package that was being used to generate the data or for the analysis, but it was unclear in the remaining 36 articles what statistical package was used to conduct the simulations. The relationship between generated samples was rarely stated within published simulation studies. Only one article stated that the simulations started with different seeds [32], whilst two other articles reported that independent samples were generated but did not explicitly mention anything about the starting seeds.

The number of simulations performed varied from 100 to 100 000 replications, with the most common choices being 1000 (19 articles) and 10 000 (12 articles) replications. It was unclear in four articles how many simulations were performed. Only six of these 58 articles provided any justification for the number of simulations performed. Three articles based their justifications on the expected SE given the number of simulations [33–35]. Two articles provided a justification in terms of the power to detect differences of a specified level from the true value as statistically significant [36, 37]. The last considered the chosen number of simulations to be sufficient, as they were not aiming to estimate any quantities with high accuracy [38].

The distributions and parameter specifications for generating the data were based on a real data set in eight of the simulation studies. In a further 16 articles, the simulated data intended to be typical of real data, although not explicitly based on a particular data set. The remaining 34 articles had no clear justification for the choices of parameters for the specified models used to generate the simulated data sets.

Generally the results from only a small proportion of the scenarios investigated were reported in an article, probably due to the limited space available. The choice of results to publish is fairly

Table II. Summary of results from review of 58 articles.

Criteria	Frequency
Random number generator	
drand48 on the Unix/LINUX system	1
Not stated	57
Statistical Software used for analysis	
Splus	6
SAS	6
R	3
STATA	1
Mathematica	1
BUGS	1
MLWIN	1
MATLAB	1
Standalone package	2
Not stated	36
Dependence of samples/starting seed	
Samples independent	2
Different seeds used	1
Not stated	55
Number of simulations	
100	6
200	3
400	1
500	8
1000	19
5000	2
10 000	13
50 000	1
100 000	1
Unclear	4
Any justification for number of simulations	
Yes	6
No	52
Justification for data generation	
Based on a real data set	8
Typical of real data	16
Not stated	34

arbitrary and can depend on the important conclusions to be portrayed. However, one article has made available the full set of simulation results, which can be downloaded from a website specified in the article [3].

4. DISCUSSION

The advances in computer technology have allowed simulation studies to be more accessible. However, performing simulations is not simple. In any simulation study, many decision are required

prior to the commencement of simulations, but there is generally no single correct answer. The choices made at each stage of the simulation process are open to criticism if not supplemented with thorough justifications.

Monte Carlo methods encompass any technique of statistical sampling employed to give approximate solutions to quantitative problems. They include, in addition to simulations, the Monte Carlo Markov chain methods such as Gibbs sampling, which are explicitly used for solving complicated integrals [39, 40]. This paper discusses simulation studies where data sets are formulated to imitate real data. Resampling studies [41, 42], where multiple data sets are sampled from a large real data set, require the same rigorous planning as simulation studies, differing from simulation studies only in terms of the generation of the data sets. Hence, similar considerations as discussed throughout this manuscript are relevant. Simulations are also useful in decision-making and engineering systems, where computer experiments are used to model dynamic processes in order to assess the effects over time and of varying any inputs (e.g. Reference [43]). Specific considerations for designing these studies in terms of formulating the problem, defining and designing the model and the choice of inputs and outputs have been discussed elsewhere (e.g. References [43, 44]).

This paper has discussed the important considerations when designing a simulation study. They include the choice of data to simulate and the procedures for generating the required data. Choices of distributions, parameters of any models, and covariate correlation structures used to generate the data set should be justified. Before commencing simulations, careful consideration should be given to the identification of the estimates of interest, the appropriate analysis, the methods for comparison, the criteria for evaluating these methods, the number of situations to consider, and the reporting of the results. In addition, every simulation study should have a detailed protocol, documenting the specific objectives and providing full details of how the study will be performed, analysed and reported. Modifications of the simulation processes, such as altering the number of simulations or collecting additional parameters or choices of scenarios, as a consequence of emerging data are possible, but can be time-consuming if they require all simulations to be rerun. Therefore, thorough planning at the start of any simulation study can ensure that the simulations are performed efficiently and only the necessary criteria and scenarios assessed. This paper has provided a concise reference (Figure 1) for researchers to follow when designing simulation studies.

A small review of published articles in one journal has suggested that the majority of simulation studies reported in the literature are not providing sufficient details of the simulation process to allow exact replication or clear justifications for the choices made. Future published simulation studies should include details of all the simulation procedures to enable the results to be reproduced. Using separate subheadings for the objectives, methods, results and discussion, irrespective of whether it is the main focus of the article, as in Reference [33], provides clarity and can aid interpretation. In addition, encouraging researcher to consider the suggested criteria (Figure 1) might encourage more sound and reliable simulation studies to be performed and reported with credible results.

REFERENCES

1. De Angelis D, Young GA. Bootstrap method. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: New York, 1998; 426–433.

2. Kristman V, Manno M, Cote P. Loss to follow-up in cohort studies: how much is too much? *European Journal of Epidemiology* 2004; **19**:751–760.

3. Vaeth M, Skovlund E. A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Statistics in Medicine* 2004; **23**:1781–1792.

4. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**:330–351.

5. Mooney C. Conveying truth with the artificial: using simulated data to teach statistics in social sciences. *SocInfo Journal* 1995; **1**(Part 7):1–5.

6. Hodgson T, Burke M. On simulation and the teaching of statistics. *Teaching Statistics* 2000; **22**:91–96.

7. Morgan BJT. *Elements of Simulation*. Chapman & Hall: London, U.K., 1984.

8. Ripley BD. *Stochastic Simulation*. Wiley: New York, 1987.

9. Whitehead J, Zhou YH, Stevens J, Blakey G, Price J, Leadbetter J. Bayesian decision procedures for dose-escalation based on evidence of undesirable events and therapeutic benefit. *Statistics in Medicine* 2006; **25**:37–53.

10. Halabi S, Singh B. Sample size determination for comparing several survival curves with unequal allocations. *Statistics in Medicine* 2004; **23**:1793–1815.

11. L'Ecuyer P. Random number generation. In *Handbook of Computational Statistics*, Gentle JE, Haerdle W, Mori Y (eds). Springer-Verlag: New York, 2004; 35–70.

12. Marsaglia G. Random number generators. *Journal of Modern Applied Statistical Methods* 2003; **2**:2–13.

13. Masuda N, Zimmermann F. PRNGlib: a parallel random number generator library. *Technical Report*: *TR-96-08*, Swiss Center for Scientific Computing, Switzerland, 1996.

14. Marsaglia G. *The Marsaglia Random Number CDROM with the DIEHARD Battery of Tests of Randomness*. Florida State University: Florida, U.S.A., 1995.

15. Demirtas H. Pseudo-random number generation in R for some univariate distributions. *Journal of Modern Applied Statistical Methods* 2005; **3**:300–311.

16. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychological Methods* 2002; **7**:19–40.

17. Mackenzie T, Abrahamowicz M. Marginal and hazard ratio specific random data generation: applications to semi-parametric bootstrapping. *Statistics and Computing* 2002; **12**:245–252.

18. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**:1713–1723.

19. Miloslavsky M, Keles S, van der Laan MJ, Butler S. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society Series B—Statistical Methodology* 2004; **66**:239–257.

20. Sevcikova H. Statistical simulations on parallel computers. *Journal of Computational and Graphical Statistics* 2004; **13**:886–906.

21. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological Methods* 2002; **7**:147–177.

22. Lachin JM. Sample size determination. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: New York, 1998; 4693–4704.

23. Diaz-Emparanza I. Is a small Monte Carlo analysis a good analysis? Checking the size, power and consistency of a simulation-based test. *Statistical Papers* 2002; **43**:567–577.

24. Rubinstein RY. *Simulation and the Monte Carlo Method*. Wiley: New York, 1981.

25. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: New York, 2002.

26. Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. *Psychological Methods* 2001; **6**:317–329.

27. Tang LQ, Song JW, Belin TR, Unutzer J. A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine* 2005; **24**:2111–2128.

28. Leffondré K, Abrahamowicz M, Siemiatycki J. Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates: a simulation study. *Statistics in Medicine* 2003; **22**:3781–3794.

29. Rempala GA, Looney SW. Asymptotic properties of a two sample randomized test for partially dependent data. *Journal of Statistical Planning and Inference* 2006; **136**:68–89.

30. Chen HY, Little RJA. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association* 1999; **94**:896–908.

31. Kaiser JC, Heidenreich WF. Comparing regression methods for the two-stage clonal expansion model of carcinogenesis. *Statistics in Medicine* 2004; **23**:3333–3350.

32. Kenna LA, Sheiner LB. Estimating treatment effect in the presence of non-compliance measured with error: precision and robustness of data analysis methods. *Statistics in Medicine* 2004; **23**:3561–3580.

33. Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine* 2004; **23**:1663–1682.
34. Chen YHJ, Demets DL, Lan KKG. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine* 2004; **23**:1023–1038.
35. Song JW, Belin TR. Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine* 2004; **23**:2827–2843.
36. Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine* 2004; **23**:1159–1178.
37. Klar N, Darlington G. Methods for modelling change in cluster randomization trials. *Statistics in Medicine* 2004; **23**:2341–2357.
38. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Statistics in Medicine* 2004; **23**:723–748.
39. Gilks WR, Richardson S, Spiegelhalter DJ. Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*, Gilks WR, Richardson S, Spiegelhalter DJ (eds). Chapman & Hall: London, U.K., 1996; 1–19.
40. Robert CP, Casella G. *Monte Carlo Statistical Methods*. Springer-Verlag: New York, 2004.
41. Lunneborg CE. *Data Analysis by Resampling—Concepts and Applications*. Duxborg: Australia, 2000.
42. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.
43. Santner TJ, Williams BJ, Notz WI. *The Design and Analysis of Computer Experiments*. Springer-Verlag: New York, 2003.
44. Balci O. Guidelines for successful simulation studies. In *Proceedings of the 1990 Winter Simulation Conference*, Balci O, Sadowski RP, Nance RE (eds). IEEE: Piscataway, NJ, 1990; 25–32.