

# BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email [info.bmjopen@bmj.com](mailto:info.bmjopen@bmj.com)

# BMJ Open

## An introduction to statistical simulations in health research

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2020-039921
Article Type:	Communication
Date Submitted by the Author:	22-May-2020
Complete List of Authors:	Boulesteix, Anne-Laure; Ludwig-Maximilians-Universitat Munchen, Institute for Medical Information Processing, Biometry and Epidemiology Groenwold, Rolf; LUMC Abrahamowicz, Michal; Division of Clinical Epidemiology, McGill University Health Centre Binder, Harald; University of Freiburg, Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center Briel, Matthias; University Hospital Basel, Institute for Clinical Epidemiology and Biostatistics Hornung, Roman; Ludwig-Maximilians-Universitat Munchen, Institute for Medical Information Processing, Biometry and Epidemiology Morris, Tim; MRC Clinical Trials Unit at UCL, ; Rahnenführer, Jörg; TU Dortmund, Department of Statistics Sauerbrei, Willi; University of Freiburg Hospital, Institute for Medical Biometry and Statistics
<b>Primary Subject Heading</b>:	
Secondary Subject Heading:	
Keywords:	STATISTICS & RESEARCH METHODS, EPIDEMIOLOGY, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT

SCHOLARONE™  
Manuscripts

## An introduction to statistical simulations in health research

Anne-Laure Boulesteix<sup>1</sup>, Rolf Groenwold<sup>2,3</sup>, Michal Abrahamowicz<sup>4</sup>, Harald Binder<sup>5</sup>, Matthias Briel<sup>6,7</sup>, Roman Hornung<sup>1</sup>, Tim Morris<sup>8</sup>, Jörg Rahnenführer<sup>9</sup>, Willi Sauerbrei<sup>5</sup>

on behalf of the Simulation Panel of the STRATOS initiative

\*To whom correspondence should be addressed: Anne-Laure Boulesteix, IBE, Marchioninstr. 15, 81377 Munich, Germany. [boulesteix@ibe.med.uni-muenchen.de](mailto:boulesteix@ibe.med.uni-muenchen.de), +49 89 440077598

<sup>1</sup> Institute for Medical Processing, Biometry and Epidemiology, Ludwig Maximilian University of Munich, Munich, Germany

<sup>2</sup> Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, the Netherlands

<sup>3</sup> Department of Biomedical Data Science, Leiden University Medical Centre, Leiden, the Netherlands

<sup>4</sup> Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada

<sup>5</sup> Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany

<sup>6</sup> Department of Clinical Research, Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel and University of Basel, Basel, Switzerland

<sup>7</sup> Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada

<sup>8</sup> MRC Clinical Trials Unit at UCL, London, UK

<sup>9</sup> Department of Statistics, TU Dortmund University, Dortmund, Germany

Word count: 6556

## ABSTRACT

In health research, statistical methods are frequently used to address a wide variety of research questions. For almost every analytical challenge, different methods are available. But how do we choose between different methods and how do we judge whether the chosen method is appropriate for our specific study? Like in any science, in statistics experiments can be run to find out which methods should be used under which circumstances based on empirical findings. The main objective of this paper is to demonstrate that simulation studies, i.e. experiments investigating synthetic data with known properties, are an invaluable tool for addressing these questions. We aim to provide a first introduction to simulation studies for data analysts or, more generally, for researchers involved at different levels in the analyses of health data, who (i) may rely on simulation studies published in statistical literature to choose their statistical methods and who, thus, need to understand the criteria of assessing the validity and relevance of simulation results and their interpretation; and/or (ii) need to understand the basic principles of designing statistical simulations in order to efficiently collaborate with a more experienced colleague or to start learning to conduct own simulations. We illustrate the implementation of a simulation study and the interpretation of its results through a simple example inspired by recent literature, which is completely reproducible using the R-script available from the supplement.

## ARTICLE SUMMARY: STRENGTHS AND LIMITATIONS OF THIS STUDY

- This paper provides a first introduction to simulation studies, i.e. experiments investigating synthetic data with known properties, which are an invaluable tool facilitating the choice of appropriate statistical designs and analysis methods.
- It does not provide details about complex issues related to simulation studies and is therefore less relevant for data analysts with experience in this field than for those with limited background.
- In the context of the corona crisis the public becomes more interested in data, its analysis, interpretation and consequences in terms of risk assessment, patient handling and prevention strategies: a simpler paper would be needed for such a group of readers.

## 1 INTRODUCTION

In health research, statistical methods are frequently used to address a wide variety of research questions. For almost every analytical challenge, different methods are available. But how do we choose between different methods and how do we judge whether the chosen method is appropriate for our specific study? Most statistical methods are developed under specific assumptions, but these assumptions are often difficult to check in applications. Moreover, performance of methods may still be reasonable when some assumptions are violated, such as the linearity of effects in regression models in the presence of mild non-linear effects. In real-life studies of human health, some of these formal underlying assumptions may be questionable or definitely violated. For example, frequent problems, such as unusual distributions, missing data, measurement errors, unmeasured confounders, or lack of accurate information on event times, may affect the accuracy or even the validity of the proposed analyses. What conditions (e.g., what sample size) are needed for a specific method to behave well? Which method is most appropriate in a particular setting?

The main objective of this paper is to demonstrate that simulation studies, i.e. evaluation of synthetic data with known properties, are an invaluable tool for addressing these questions. We aim to provide a first introduction to simulation studies for data analysts or, more generally, for researchers involved at different levels in the analyses of health data, who (i) may rely on simulation studies published in statistical literature to choose their statistical methods and who, thus, need to understand the criteria of assessing the validity and relevance of simulation results and their interpretation; and/or (ii) need to understand the basic principles of designing statistical simulations in order to efficiently collaborate with a more experienced colleague or to start learning to conduct own simulations. Statisticians interested in more details about statistical simulations are referred to the more technical overviews available in the literature.[1-3]

Statistical methodology has seen substantial development in recent times but many of these developments are largely ignored in the practice of health data analyses. To help bridge the gap between methodological innovation and applications, the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative was launched in 2013.[4] It aims to provide statistical guidance for key topics in the design and analysis of observational studies. In practice, analyses are sometimes conducted by researchers with limited statistical background. Consequently, STRATOS plans to develop guidance for researchers with different levels of statistical knowledge including researchers without strong statistical backgrounds (see Table 1 in [4]). For the analysis of observational studies, typically several approaches are possible, and the properties of each approach should be assessed in comparison with alternative methods. Simulation studies are key instruments for such assessments. Ideally, all data analysts should be familiar with them.

This paper is structured as follows. We first discuss the role of statistical simulation studies in section 2. Section 3 outlines four relatively simple examples of statistical methods and then

1  
2  
3 explains how the performance of these methods could be evaluated using simulation  
4 studies. Section 4 sketches out the basic principles of designing and conducting simulations.  
5 Finally, section 5 briefly illustrates the implementation of a simulation study and the  
6 interpretation of its results through a simple example inspired by recent literature.  
7  
8  
9

## 10 11 12 **2 THE ROLE OF SIMULATION STUDIES**

### 13 14 **Comparing methods based on theory**

15  
16 During the first half of the 20th century, mathematical theory was the cornerstone of  
17 evaluating traditional statistical methods addressing well defined problems. However, to  
18 investigate questions in modern medicine, more complex statistical modelling or the use of  
19 machine learning techniques are often required. Only in rare cases of low complexity and  
20 often of limited practical relevance, mathematics tells us that - given the data satisfy certain  
21 properties - the considered method behaves in a particular way. For example, theory tells us  
22 that the two-sample t-test has better power to detect a true difference between mean values  
23 in two independent groups than the Mann-Whitney test - if the variable of interest is normally  
24 distributed within each of the two groups. Most theoretical results of this type are valid only  
25 under specific assumptions about the available data. While it may be acceptable to assume  
26 normally distributed data in the case of the simple example mentioned above, for more  
27 complex problems the required assumptions can be unrealistic; see sections 3.2, 3.3 and 3.4  
28 for examples beyond this simple case. Moreover, the process of verifying assumptions is  
29 often already challenging in practice; see for example [5] for an extensive simulation study  
30 on the choice between t-test and Mann-Whitney-test including considerations on normality  
31 checks.  
32  
33  
34  
35  
36  
37  
38

### 39 40 **Comparing methods using empirical data**

41  
42 Another approach for evaluating statistical methods consists of applying them to  
43 representative datasets from the considered field and assessing their performance; or, more  
44 generally, of observing their behaviour when using them in these datasets. Some important  
45 characteristics of statistical methods can indeed be derived from real datasets. For example,  
46 are results stable if we modify the dataset slightly? For many approaches, however, the most  
47 important evaluation criteria cannot be assessed for real data, simply because for real data  
48 we do not know the true values of the underlying parameters we aim to draw inferences  
49 about. For example, if one method estimates a difference of 1 between two groups, and  
50 another estimates a difference of 2, we can see that they give us different results (assuming  
51 that the confidence intervals are narrow), but we do not know whether 1 or 2 is closer to the  
52 correct answer.  
53  
54  
55  
56  
57  
58  
59  
60

### Why simulation studies?

A simulation study is useful if theoretical arguments are insufficient to determine whether the method of interest is valid in a specific real-life application or whether violations of the assumptions underlying the available theory (such as large sample size, normal distribution of residuals, proportional hazards, etc.) affect the validity of the results. In methodological research, simulations play a role similar to experiments in basic science.[6] The idea of a simulation study is to investigate the behaviour of methods when applied to synthetic datasets with known characteristics. Because the 'correct' or 'true' answer is known by the researchers, who had full control of how the data were simulated, simulations permit assessing whether the methods recover this known truth. For example, we may generate data with and without a treatment effect and then assess how often a test correctly or incorrectly rejects the null hypothesis of no treatment effect. Alternatively, we may generate data in which the treatment effect has a certain value and then study how accurately a regression model can estimate this known effect. Notice that such assessment is *not* possible using real data when the true response or the true effect is not known.

Suppose a scientist is planning a cohort study of the effect of an exposure on time to a clinical event (e.g. death) and wants to know what sample size is necessary to achieve a certain power with a given test, or a certain precision with a given estimation method. A question that might be explored using a simulation study could be the following: What is the power of the logrank-test (an asymptotic test requiring large sample sizes to ensure validity), in the case of small samples? Here, a simple simulation study, designed to be consistent with the specific settings of the proposed study (sample size, prevalence of the exposure of interest, incidence of events, etc.), could provide the necessary answers.

Simulation studies are also helpful to provide objective reproducible answers to more general methodological questions on the behaviour of statistical methods (i.e., not necessarily motivated through a specific application). Examples of this type of question, which have been investigated by recent simulation studies, include: What is the effect of measurement errors on the estimated exposure-outcome relations in epidemiological studies?[7] Does it make sense to check for subgroup-specific treatment effects even if the test for an overall effect is non-significant?[8]

In addition to the evaluation of individual methods, simulations can also be used to determine which one of several candidate methods will perform best for the application at hand. In the case of simulations reported in statistical literature, candidate methods may include existing methods, and may (but do not have to) include new methods proposed by the researchers performing the simulation study. In the latter case, their focus is often on showing in which settings the new method performs better than its existing 'competitors'. [9, 10]

No matter the context of the simulation study, the objective is to find out if/when methods perform well or break. Regarding the "when" question, simulations provide an ideal setting

1  
2  
3 for a systematic assessment of how variations in the values of relevant parameters, and/or  
4 assumptions regarding data structure (e.g. independence of observations, lack of  
5 measurement errors) affect the performance of the methods of interest. The definition of the  
6 term “good performance” depends on the context. For example, if we compute a 95%-  
7 confidence interval, we usually want it to yield 95% coverage (i.e., we want 95% of the  
8 confidence intervals constructed in this way, using varying datasets, to cover the true value).  
9  
10 If we apply a statistical test, we want this test to reject the null hypothesis with high  
11 probability if it is false, but to *retain* it with high probability if it is true. In comparison studies,  
12 two or more methods may be compared in this respect. In the case of a simulation  
13 performed for sample size calculation, we want to determine the smallest sample size with  
14 which a study has a given power to detect clinically important effects.  
15  
16  
17  
18

19 In practice, nobody can predict with certainty whether a method will yield accurate results for  
20 a specific dataset, or which of a set of considered methods will perform best on that dataset.  
21 Simulations can provide *systematic evidence* regarding how methods perform on average for  
22 datasets with similar characteristics to the dataset under investigation. In an ideal world,  
23 relevant results from simulation studies would be available from previous research to help  
24 make rational decisions about which method to use. Data analysts would then use simulation  
25 results to verify whether the method they choose is adequate, or to pick the most suitable  
26 from a range of different methods. Such “previous research” is typically done by statistical  
27 researchers working on methods as the focus of research (as opposed to researchers  
28 *applying* methods in health research projects). For a data analyst with little experience and  
29 background in statistical methodological research, it is important to be able to interpret the  
30 results of such simulation studies. If previous evidence is lacking, or if previous studies do  
31 not seem to apply to the specific data setting under consideration, data analysts should  
32 conduct a targeted simulation study tailored to their specific dataset.  
33  
34  
35  
36  
37  
38  
39  
40  
41

### 42 **3 EXAMPLES OF STATISTICAL METHODS**

43  
44 In this section we present four examples of analyses which help us explain the basic  
45 principles of simulation studies. Key criteria for evaluating the performance of methods  
46 related to these examples are summarised in Table 1, at the end of the section.  
47  
48

#### 49 **3.1 Statistical hypothesis testing and confidence intervals**

50  
51 In most health research projects we perform statistical tests and/or derive confidence  
52 intervals. However, their behaviour is often not well-characterised in real world situations.  
53 For example, for time-to-event data with censored observations, how do the logrank-test and  
54 confidence intervals for the hazard ratio behave in small sample settings? Which technique  
55 should be preferred to compute confidence intervals for proportions in a given setting (e.g.,  
56 very small proportions)?[11]  
57  
58  
59  
60



- What is a good test/confidence interval?

A good test is one that yields the correct answer with high probability, i.e. one that rejects the null hypothesis with high probability if it is not true, and retains it with high probability if it is true. Classical tests are defined in such a way that, in theory, the probability that the null hypothesis is rejected despite being true (called type 1 error) does not exceed a level  $\alpha$  chosen by the user (in medicine, often  $\alpha=0.05$ ) - provided the assumptions are fulfilled. However, it is possible that the actual type 1 error may be larger than  $\alpha$ , in which case the results of the test should be interpreted with caution. When evaluating a test, it is thus important to verify that the type 1 error does not exceed the nominal significance level  $\alpha$  that was chosen by the researcher. Provided the type 1 error is as it should be (equal to or smaller than  $\alpha$ ), the most important quantity characterising a statistical test is its power, defined as the probability of correctly rejecting the null hypothesis.

Apart from hypothesis testing, results of statistical analysis are oftentimes presented as an estimate with a corresponding confidence interval. A good method for deriving, say, 95% confidence intervals is a method that yields confidence intervals covering the true value with probability 95%.

- Can real data be used for the evaluation?

The main performance criteria cannot simply be assessed based on real data, because the truth (which hypotheses are true or false, or the true value of the parameter to estimate) is generally unknown in practice - we can see that a test has rejected the null hypothesis, but do not know if this was correct or not. If the truth were known, there would be no need to perform the test or compute a confidence interval. Baseline characteristics in correctly randomised trials are a notable exception. Given the randomisation procedure, they are expected to be equally distributed in the two groups by definition.

### **3.2 Model selection for regression models: explaining the effects of independent variables on a dependent variable**

The second example is regression modelling of a dependent variable of interest (typically, a clinical outcome) using several independent variables (often, prognostic, or risk factors). In general, such modelling is performed either to *explain* the dependent variable by determining the effects of the independent variables (as considered in this section), or to build a model, which will be used later on new patients for *prediction* purposes (as considered in the next section); see [12] for a discussion of these two related but distinct purposes. In health research, the dependent variable is often of one of the three following types: continuous (e.g., amount of cholesterol reduction), categorical (e.g., response to therapy) or survival time (e.g., disease free survival in months). Even though for all three cases standard regression modelling is reasonably well-understood, the behaviour of regression techniques

1  
2  
3 (including model selection) still raises questions in particular cases; see for example a recent  
4 simulation study on the use of resampling techniques for model selection purposes.[13]  
5  
6

- 7 • What is a good regression approach?

8  
9 In principle, a regression technique (including model selection aspects) is expected to (i)  
10 correctly distinguish the variables that have an effect on the dependent variable from those  
11 that have no effect, and (ii) correctly fit the regression coefficients of the variables, i.e. fit  
12 them to provide estimated values close to the true ones (unbiased and low variance).  
13 Regarding (i), it is good to have high sensitivity (i.e., selecting all variables with effects) as  
14 well as high specificity (i.e., not selecting variables without an effect). Depending on the  
15 specific aim analysts may also aim to eliminate variables with very small effects.  
16  
17  
18

- 19 • Can real data be used for the evaluation?

20  
21 In practice, the exact set of variables that have an effect on the dependent variable and the  
22 values of these effects are unknown, although previous knowledge from the literature may  
23 provide valuable guidance in some cases. Thus, in most cases, real data are of limited use  
24 for the evaluation of model selection approaches for regression models.  
25  
26  
27

### 28 **3.3 Model selection for regression models: predicting the values of an outcome using** 29 **the values of independent variables**

30  
31 The third example is related to the second example, but takes a different perspective. While  
32 regression models are often used to “explain” the dependent variable (e.g., a disease  
33 outcome or survival time), in order to understand how different risk factors affect the  
34 dependent variable, they can also be used as “prediction models” to predict the outcome of  
35 interest (also called “dependent variable”) for new patients, based on these patients’ values  
36 of the predictor variables (also called “covariates” or “independent variables”). Classical  
37 linear regression models can be used for this purpose as well as various more complex  
38 alternative procedures, especially algorithms developed in the machine learning community,  
39 such as support vector machines or random forests (see [14] for a gentle introduction). In  
40 this field, simulations can be useful to assess the prediction accuracy of the considered  
41 prediction methods in different settings. For example, different penalised regression methods  
42 may be compared in simulations with respect to their prediction performance when a small  
43 number of clinical covariates are combined with a large number of candidate molecular  
44 covariates.[15]  
45  
46  
47  
48  
49  
50  
51

- 52 • What is a good prediction model?

53  
54 A good prediction model is a model that yields accurate predictions in the future patients it  
55 will be applied to. For continuous and for categorical dependent variables, often predicted  
56 and true values are directly compared, and the differences are summarised across patients.  
57  
58  
59  
60

1  
2  
3 For survival times, suitable adjusted scores, like the Brier score, may be used to take into  
4 account censoring.[16]  
5  
6  
7  
8

- 9
- Can real data be used for the evaluation?
- 10

11 The prediction error can be estimated based on the available dataset using a large (possibly  
12 external) validation dataset if available, or so-called resampling techniques such as cross-  
13 validation.[17] Note that this estimation may be unreliable depending on the context (for  
14 example, the smaller the sample size, the more unstable the cross-validation estimates).[18]  
15 What these evaluations tell us about the methods' accuracy is relevant to the considered  
16 specific real data example(s) but may not be relevant to other settings.  
17  
18  
19

### 20 21 **3.4 Clustering**

22  
23 The last example considered in this paper is clustering, also called cluster analysis. The  
24 objective of clustering is to identify clusters, i.e., “groups” of patients that behave similarly.  
25 For example, clustering methods may be used with the goal of identifying clinically  
26 meaningful subgroups of patients, using magnetic resonance imaging data and clinical data,  
27 among others.[19] Clusters should be constructed in such a way that the values of patients  
28 within a cluster are more similar (according to the chosen similarity criterion) than values of  
29 patients from different clusters. Many different clustering algorithms have been proposed at  
30 the interface between computer science and statistics, for example k-means clustering or  
31 hierarchical clustering. Simulation studies may be used to assess the ability of methods to  
32 recover a true underlying structure.[19, 20]  
33  
34  
35  
36  
37

- What is a good clustering method?
- 38  
39

40 A good clustering procedure is a procedure that correctly recovers a true cluster structure  
41 present in the data.  
42  
43

- Can real data be used for the evaluation?
- 44  
45

46 In practice, the true cluster structure is often unknown. And even if there is a known cluster  
47 structure, further sensible cluster structures might exist. The abilities of clustering methods to  
48 group similar observations together may be assessed by using data that consists of known  
49 subgroups and measuring the degree of overlap between the clustering structure defined by  
50 the known subgroups and the clustering structure proposed by the clustering algorithm;  
51 however, there might not be only one sensible cluster structure; in fact, the observations may  
52 cluster together more strongly according to other factors than the subgroup membership,  
53 e.g., gene expressions are associated with various phenotypes. Real data may be used to  
54 assess aspects such as stability or computational efficiency, but they are of limited use for  
55 the evaluation of a clustering method according to the criterion “agreement with the true  
56 cluster structure”.  
57  
58  
59  
60

Example	evaluation criterion	aim
<b>A – testing and confidence intervals</b>	type 1 error type 2 error coverage	low low close to nominal value
<b>B – explaining</b>	mean coefficient values precision of coefficient estimation coverage sensitivity of variable selection specificity of variable selection	close to true ones (low bias) high (low variance) close to nominal value high high
<b>C – predicting</b>	prediction error on independent data accuracy measures	low high
<b>D – clustering</b>	agreement with true cluster structure	high
<b>A-B-C-D</b>	stability computational cost model convergence interpretability	high low achieved high

Table 1. Overview of the main criteria for evaluating statistical methods in the four considered examples.

## 4 BASIC PRINCIPLES OF SIMULATION STUDIES

### 4.1 Key features of a simulation study

In this section we give a brief overview of the key features of a simulation study. A more detailed introduction to the concepts of data generating mechanisms and simulation

1  
2  
3 scenarios is given in section 4.2, for interested readers. One may also refer to a recent in-  
4 depth article on simulation studies addressing an audience of statisticians.[3]  
5

6  
7 The first key feature of a simulation study is its *overall objective*. Is the simulation study  
8 tailored to a specific dataset relevant for a particular application or does it address a  
9 methodological question of general interest for future applications? Regardless of the overall  
10 objective, researchers performing a simulation study should make decisions considering the  
11 following key issues.  
12  
13

14  
15 *Choice of methods to be evaluated/compared*: Which method(s)/variant(s) is (are)  
16 evaluated? This point is analogous to the definition of the treatments with all necessary  
17 details (dose, etc.) to be compared in a clinical trial. Further discussion about the analogy  
18 between clinical trials and comparisons of statistical methods can be found elsewhere.[9]  
19

20  
21 *Specific aims*: What do we want to learn about the method(s) from the simulation study? For  
22 example, one may want to assess whether a model selection method selects the right  
23 covariates (main aim), and whether it estimates their effects accurately (secondary aim).  
24 This point is analogous to the definition of primary and secondary outcomes in clinical trials,  
25 e.g., disease-free survival or side effects.  
26  
27

28  
29 *Data generating mechanism (including choice of relevant parameters)*: How do we generate  
30 the simulated datasets? From which distribution? Which parameters may affect the results  
31 and what values should be considered? Each combination of the relevant assumptions and  
32 parameter values defines one simulation scenario (for which several datasets will usually be  
33 (randomly) generated, as outlined in the next section). There are many ways to generate  
34 datasets: using real datasets as a basis or by sampling from (possibly multivariate) pre-  
35 specified distributions, e.g., the normal distribution. The definition of the scenarios is  
36 analogous to the definition of experimental conditions for a lab experiment, and should be  
37 guided by considerations about clinical plausibility and/or relevance.[10] While simulation  
38 designs can be made arbitrarily complex, the focus is often on relatively simple properties of  
39 the data distributions, such as skewness or outliers. The performance of many widely used  
40 basic statistical building blocks, such as the least squares optimisation principle for  
41 estimating model parameters, can be severely affected by the type of distribution under  
42 consideration. As a result, in order to comprehensively gauge performance, simulation  
43 studies should also include the rather innocent looking problems of real data, such as some  
44 outlier observations. More insights are given in section 4.2.  
45  
46  
47  
48  
49  
50

51  
52 *Performance measure(s)*: Which criteria are used to assess the performance of the  
53 considered data analysis methods? In the example of model selection mentioned above, one  
54 may address the main aim by considering the sensitivity of the method for selecting the “true  
55 effects” as well as the frequency of “false positives” (i.e. selection of variables that have no  
56 true associations with the outcome). The secondary aim may be addressed by computing  
57 the mean squared deviation or the mean absolute deviation of the coefficient estimates from  
58 the true values. This point is analogous to the precise definition of primary and secondary  
59  
60

1  
2  
3 outcomes in a clinical trial: e.g., which instruments are used for the assessment of side  
4 effects of the therapy, or how do we exactly estimate disease-free survival and compare it  
5 across the trial arms?  
6  
7

8 *Number of repetitions:* For each considered scenario, how many datasets are randomly  
9 drawn? It is necessary to generate several (ideally, “many”) datasets in order to average out  
10 random fluctuations and ensure sufficiently precise simulation results. The more datasets are  
11 generated, the more precise the performance evaluation will be - as can be quantified  
12 through, for example, the width of the confidence intervals for the selected “performance  
13 criteria”. The number of repetitions is analogous to the sample size in a clinical trial. In  
14 contrast to increasing the sample size in clinical trials, however, it is often easy to extend the  
15 number of repetitions in simulation studies. The number of repetitions is chosen as a  
16 compromise between precision of the results and computational time.  
17  
18  
19  
20  
21

## 22 **4.2 Sampling variability and data generating processes**

23  
24 This section gives further insights into the data generating process for readers interested in  
25 gaining a deeper understanding of the fundamentals of simulation studies, beyond the key  
26 points outlined above. To this end we first explain briefly how simulations provide a  
27 framework for assessing and accounting for the impact of random sampling error on the  
28 results of empirical studies. Suppose a clinical researcher is interested in the mean  
29 difference between the blood pressure of males and females in the population aged 20 to 60.  
30 The true mean difference could only be calculated if we had data on the whole populations of  
31 males and females aged 20 to 60. Of course, in practice, we only have a sample available  
32 with a specific (often moderate) size and can only *estimate* the mean difference using this  
33 sample. Different samples will yield different estimates of the same mean difference in the  
34 population. Collecting a data sample can be seen as drawing observations from a population  
35 of interest that has particular characteristics. In statistical terms, these observations can be  
36 seen as random observations generated from the *true distribution of the variable(s) of*  
37 *interest in* the relevant population. In real-life studies, this distribution and its true parameters  
38 (e.g., population means) are unknown and we can only *estimate* them using available  
39 sample data.  
40  
41  
42  
43  
44  
45  
46

47 The principle of simulations is to mimic the process of taking random samples from a large  
48 population, by repeatedly generating synthetic data (“virtual observations”) from a virtual  
49 population, under pre-specified assumptions that can be varied across the considered  
50 simulation scenarios. Each synthetic sample is generated from a particular known  
51 distribution, with “true” values of all relevant parameters fixed by the researchers. Each  
52 simulated sample is then analysed using the method(s) of interest, and its (their)  
53 performance is evaluated using pre-specified criteria (see Table 1 for examples). To give  
54 one simple example, we may simulate systolic blood pressure (SBP) values for a sample of  
55 N=100 “synthetic subjects” by generating 100 independent numbers from a normal  
56 distribution with, say, mean 120 and standard deviation 15. Doing so, we know that the true  
57  
58  
59  
60

1  
2  
3 population mean is 120 mm Hg and that the simulated blood pressure follows the normal  
4 distribution. The way in which virtual observations are generated in the context of a  
5 simulation (in our example, “100 independent numbers from a normal distribution with mean  
6 120 and standard deviation 15”) is termed the *data generating mechanism*. There is a large  
7 number of user-friendly statistical packages that can be used to accomplish this task.  
8  
9

10  
11 Just as random sample-to-sample variability affects real data samples drawn from a  
12 population of interest, it also affects the results obtained using simulated data. If we generate  
13 two synthetic datasets using the same data generating mechanism and the same  
14 parameters, we will get somewhat different results (with the differences decreasing, on  
15 average, with increasing size of the generated datasets). It is therefore almost always  
16 important to repeat the same data generation and analysis process using many simulated  
17 datasets, as outlined in the paragraph “*Number of repetitions*” of section 4.1 above. The  
18 variability of the results obtained across the different datasets simulated from the same  
19 distribution has to be carefully assessed by, for example, calculating the standard deviation  
20 of the individual estimates. On the other hand, calculating the mean value of the individual  
21 estimates provides a more robust estimate of the unknown population-level parameter than a  
22 value from a single simulated sample, as averaging over several repetitions reduces the  
23 impact of random sampling error.  
24  
25  
26  
27  
28

29  
30 When performing a simulation, one has to choose one or several data generating  
31 mechanisms that reflect, as closely as possible, the distribution and relevant characteristics  
32 of the real data of interest, no matter whether the focus is on a specific application or on a  
33 ‘generic’ methodological question such as evaluation or comparison of specific analytical  
34 methods. The difficulty is that, in reality, the true data generating process is unknown as  
35 mentioned above in the example of blood pressure. The only possibility is to consider  
36 several data generating mechanisms – called simulation scenarios – that, together, will cover  
37 the range of situations congruent with the expected structure of real data of interest. For  
38 example, we may be interested in the behaviour of a test, that assumes a normal  
39 distribution, in situations where this assumption is not fulfilled. If the variable of interest is  
40 expected, based on earlier studies and/or substantive knowledge, to be (approximately)  
41 uniformly distributed (meaning that the observations are evenly distributed over a certain  
42 interval), priority will be given to corresponding scenarios. However, it may be useful to also  
43 consider a few alternative scenarios with other distributions, e.g., a positively skewed  
44 distribution with most values concentrating below the mean and relatively fewer high  
45 values.  
46  
47  
48  
49  
50  
51

52  
53 In general, if the focus of the simulation study is on a specific application, the primary goal is  
54 essentially to simulate datasets that are as similar as possible to the relevant real dataset.  
55 This may necessitate making some plausible assumptions and involve some uncertainty if  
56 the data have not yet been collected – as is the case when simulations are performed with  
57 the aim of calculating the adequate sample size or assessing the expected power and/or  
58 precision of future analyses. In contrast, if the focus of the simulation is on the general  
59  
60

1  
2  
3 behaviour of a particular method (or comparison of alternative methods) for a class of  
4 applications, the primary goal when choosing scenarios is often to cover a wide spectrum of  
5 potentially plausible situations, in which the method(s) of interest are likely to be employed.  
6 Some scenarios may be unrealistic but are nevertheless helpful in understanding how the  
7 method works or when it breaks down (and how it can be improved to cope better with the  
8 problematic situations), and thus yield valuable information. The choice of simulation  
9 scenarios is thus intrinsically related to the goal of the simulation, but should also account for  
10 substantive knowledge in the field of potential real-life applications.  
11  
12  
13

#### 14 15 **4.3 Advantages and drawbacks of simulation studies**

16  
17 To simulate the synthetic datasets, we define the underlying “truth” regarding the research  
18 question being explored. For example, in example A in section 3 (testing) we know whether  
19 the null hypothesis is true or not. In example B (explaining) we know which variables have  
20 independent effects on the dependent variable. In example C (predicting) we know the true  
21 values of the dependent variable. In example D (clustering) we know the true cluster  
22 structure. To sum up, in all these examples, we know what an *accurate* method of data  
23 analysis is supposed to find. Thus, we can determine how well the method(s) being  
24 evaluated perform(s) by comparing their results against this known “truth”. This feature is the  
25 major advantage of simulations over empirical comparisons of the same methods based on  
26 one or few real-life datasets as, in the latter case, the true answers often remain unknown.  
27  
28  
29  
30

31  
32 Another advantage of simulations is that they allow investigation of a large number of  
33 different scenarios, and in particular also scenarios that are not directly observed in real  
34 datasets. This means that the analysis can be extended to new or rare scenarios, or  
35 scenarios reflecting ethically unacceptable or practically unrealistic settings (e.g.,  
36 randomisation or very large sample sizes). A related advantage of simulations is that, by  
37 varying the assumptions and the values of relevant parameters used to generate data for  
38 different scenarios, one can *systematically* assess how the performance of different methods  
39 depends on these assumptions and parameters. Furthermore, one can also perform, for  
40 each considered scenario, as many repetitions as needed to average out random  
41 fluctuations. This is in contrast to real data experiments where the quantity of data is often  
42 severely limited, which affects the precision of the results.  
43  
44  
45  
46  
47

48 These advantages, however, come at a cost. Firstly, simulation scenarios are often  
49 simplified, i.e. do not reflect the true complexity of the data encountered in real-life data  
50 analyses. The lack of complexity of simulated data may lead to a distorted picture of the  
51 methods' performances. For example, an approach that can model data in a very flexible  
52 manner might be more severely affected by outliers. Yet, simulation designs so far rarely  
53 incorporate outliers or skewed distributions. Real-world performance of an approach that has  
54 been selected based on simulation study results might be surprisingly bad. Secondly, large  
55 simulation studies can be computationally very expensive, taking days or weeks and even  
56 requiring the use of parallel computing, if a large number of scenarios and/or large numbers  
57  
58  
59  
60



1  
2  
3 of repetitions are considered and especially if the analysis also involves large datasets  
4 and/or complex statistical methods.  
5  
6

7 Finally, it is important to note that simulations are not immune to the typical flaws of  
8 numerical studies leading to biased results. The effect of single influential points, which are  
9 difficult to detect in simulation studies with hundreds or thousands of simulated samples, can  
10 be critical. They may be relevant in some of the simulation repetitions, in which they cause  
11 unreliable results. If undetected, they can bias the results. Most importantly, selective  
12 reporting may be an issue. If a very large number of scenarios is analysed, but only those  
13 scenarios that favour one particular method are presented in the paper, the reported results  
14 will give a distorted picture of reality. Obviously, this is a serious problem of bad reporting  
15 and bad research, which can be easily avoided by being honest.  
16  
17  
18  
19  
20  
21

## 22 **5 AN EXAMPLE OF A STATISTICAL SIMULATION**

23  
24  
25 For illustration, in this section we consider a simple simulation study that investigates the  
26 impact of measurement error in linear regression analysis, inspired by a previous study.[7]  
27 Our study is completely reproducible using the R code provided in the supplement, which  
28 uses freely available data. In epidemiological studies of the relation between an exposure  
29 and an outcome, this relation is often estimated using regression analysis. As an example,  
30 we consider a study of the association between glycosylated haemoglobin levels (HbA1c) and  
31 systolic blood pressure assessed using linear regression. Data from 5092 subjects in the  
32 2015–2016 National Health and Nutrition Examination Survey (NHANES)[21] is used to  
33 obtain an estimate of the effect of HbA1c on systolic blood pressure, while adjusting for age,  
34 gender, and body mass index (BMI). Details on the data are described on the NHANES  
35 website: [<https://wwwn.cdc.gov/nchs/nhanes/>]. After adjustment for age and gender, it was  
36 estimated that HbA1c increases systolic blood pressure by 1.13 mmHg (95%CI 0.73 – 1.52)  
37 per unit increase in HbA1c. Additional adjustment for BMI resulted in a considerable change  
38 in the effect estimate: HbA1c was estimated to increase blood pressure by 0.75 mmHg  
39 (95%CI 0.35 – 1.16) per unit increase in HbA1c.  
40  
41  
42  
43  
44  
45

46 The confounding variable BMI as well as the exposure variable HbA1c may be subject to  
47 measurement error. For example, BMI may be self-reported (instead of a standardised  
48 measurement using scales) or technical problems in the lab may have affected the HbA1c  
49 measurement. Therefore, researchers may want to know the possible impact of  
50 measurement error of the exposure and/or confounding variable(s) in terms of bias.[22] We  
51 are interested both in the direction and magnitude of this bias.  
52  
53  
54  
55

56 One way to investigate the possible impact of measurement error is through a small  
57 simulation study.[7] For the purpose of this example, the original recordings in the NHANES  
58 data were assumed to be measured without error. Then, in addition, new artificial variables  
59 were created that represented HbA1c and BMI, but for the situation in which these are  
60

1  
2  
3 measured with error. To create these variables, measurement error was artificially added to  
4 the exposure variable (HbA1c) and/or the confounding variable (BMI). These errors were  
5 drawn from a normal distribution with mean zero, and were independent of all variables  
6 considered. This type of measurement error is often referred to as classical measurement  
7 error.[23] The variance of the normal distribution was varied in different scenarios, in order to  
8 reflect varying amounts of measurement error. Scenarios ranged from no measurement error  
9 on either HbA1c or BMI (reference scenario) to 50% of the variance in HbA1c and/or BMI  
10 attributable to measurement error. To minimise the impact of simulation error, each scenario  
11 was repeated 1,000 times and results were averaged per scenario over these 1,000  
12 repetitions.  
13  
14  
15  
16

17  
18 Figure 1 shows the impact of measurement error on HbA1c and/or BMI. The relation  
19 between HbA1c and systolic blood pressure was attenuated when measurement error was  
20 added to HbA1c, but not when measurement error was added to BMI. However, the  
21 association became stronger as measurement error was added solely to the confounding  
22 variable BMI. The reason for this effect is that, with increasing levels of measurement error  
23 on BMI, adjustment for the confounding due to BMI becomes less efficient and the effect  
24 estimate gets closer to the unadjusted estimate (1.13mmHg). Due to measurement error, a  
25 type of residual confounding is introduced. In the case of measurement error on HbA1c as  
26 well as BMI, both phenomena play a role and may cancel each other out. In this study,  
27 measurement error on HbA1c seemed more influential than measurement error on BMI.  
28  
29  
30  
31

32  
33 This example illustrates how a simple simulation study could provide insight into an  
34 important potential source of bias, namely measurement error. Here, we only considered  
35 classical measurement error, but simulations could easily be extended to incorporate more  
36 complex forms of measurement error. For example, the errors may not be drawn from a  
37 normal distribution with mean zero or may not be independent of all other variables  
38 considered. Instead, the mean of the distribution of errors may depend on the value of  
39 another variable in the model, e.g., error on BMI may depend on gender. Furthermore, non-  
40 normal distributions may be considered, or scenarios in which the variance of the errors  
41 depends on the true value of the measurement (heteroskedastic errors), among other  
42 possible extensions.  
43  
44  
45  
46

47  
48 Finally, we note that researchers conducting small-scale simulation studies like the one  
49 presented here should reflect on the plausibility of the scenarios considered. For example,  
50 knowing whether it is realistic to assume that 50% of the total variance of HbA1c and BMI is  
51 due to measurement error (top right scenario in Figure 1) requires subject-matter knowledge.  
52  
53  
54

## 55 **6 CONCLUDING REMARKS**

56  
57 Just as randomised clinical trials form part of the evidence base for the choice of therapy in  
58 medical practice, simulation studies form part of the evidence base for statistical practice.  
59 Large-scale simulation studies allow assessment of the properties of complex estimation and  
60

1  
2  
3 inferential methods, and comparison of complex model building strategies under a variety of  
4 alternative assumptions and sample sizes.[4] They provide valuable support for decision-  
5 making regarding the choice of statistical methods to be used in a given real-life application  
6 and they are the cornerstone of the work on guidance for the design and analysis of the  
7 STRATOS initiative. They complement - rather than replace - the judgement of a trained  
8 expert (a data analyst in the case of statistical methods, and a physician in the case of  
9 therapies). Increasing computational power nowadays makes it possible to examine many  
10 possible simulation scenarios with different combinations of distributional parameters and  
11 assumptions. This partly addresses the main limitation of simulations, namely that they can  
12 never fully reflect the complexity of real data.  
13  
14  
15  
16  
17

18 Let us again consider our analogy between simulation studies and clinical studies. The  
19 design and implementation of clinical studies should be left to teams of trained clinical  
20 researchers, but it is crucial for practitioners who want to practice evidence-based medicine  
21 to be able to read and understand the results of these clinical studies. Similarly, the design,  
22 implementation and reporting of complex simulations is still a subject of debate [3] and  
23 should be left to methodological experts, but it is important for data analysts to be able to  
24 read and understand simulation studies in the literature (or perhaps to implement simple  
25 ones themselves). Armed with these competence and analytical skills, they will be better  
26 able to identify appropriate data analysis methods for their data and research questions.  
27  
28  
29  
30  
31  
32

### 33 **ACKNOWLEDGMENTS**

34  
35 The authors thank Alethea Charlton for language corrections. The international  
36 STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative aims to  
37 provide accessible and accurate guidance for relevant topics in the design and analysis of  
38 observational studies (<http://stratos-initiative.org>). Members of simulation panel at the time of  
39 first submission: Michal Abrahamowicz, Harald Binder, Anne-Laure Boulesteix, Rolf  
40 Groenwold, Victor Kipnis, Tim Morris, Jessica Myers Franklin, Willi Sauerbrei, Pamela Shaw,  
41 Ewout Steyerberg, Ingeborg Waernbaum.  
42  
43  
44  
45  
46  
47  
48

### 49 **COMPETING INTERESTS**

50  
51 The authors declare no competing interests.  
52  
53  
54  
55

### 56 **PATIENT AND PUBLIC INVOLVEMENT STATEMENT**

57  
58 Patients or the public WERE NOT involved in the design, or conduct, or reporting, or  
59 dissemination plans of our research.  
60

## AUTHOR CONTRIBUTIONS

ALB and WS initiated and coordinated the project. ALB wrote most of the manuscript. RG performed the example analysis and wrote the corresponding section. All authors made substantial contributions to the manuscript's content, text and approved the final version.

## FUNDING STATEMENT

This project was partly funded by the German Research Foundation (DFG) with grants BO3139/4-3 to ALB and SA580/10-1 to WS. MA is a James McGill Professor at McGill University. His research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant 228203 and the Canadian Institutes of Health Research (CIHR) grant PJT-148946.

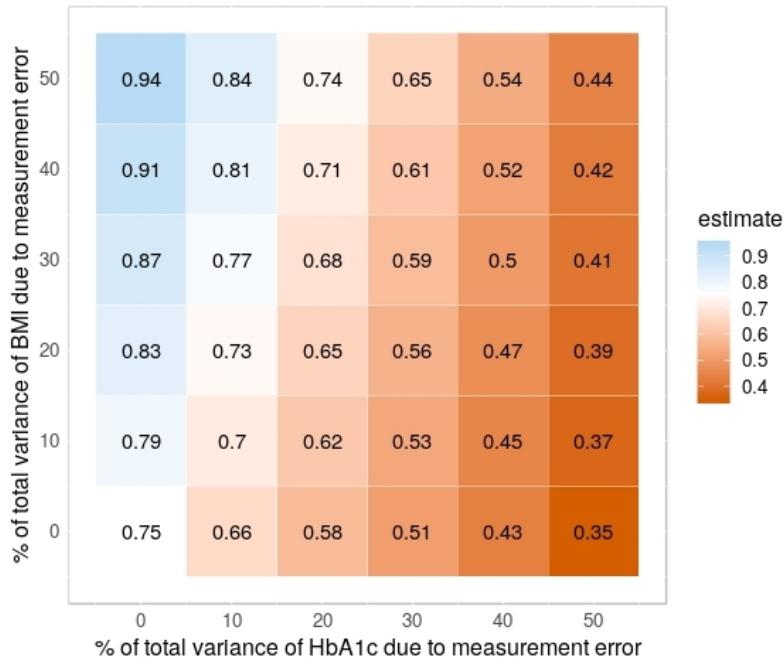
## REFERENCES

- 1 Burton A, Altman DG, Royston P, et al. The design of simulation studies in medical statistics. *Stat Med* 2006;25(24):4279-92.
- 2 Sigal MJ, Chalmers RP. Play It Again: Teaching Statistics With Monte Carlo Simulation. *Journal of Statistics Education* 2016;24(3):136-56.
- 3 Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med* 2019;38(11):2074-102.
- 4 Sauerbrei W, Abrahamowicz M, Altman DG, et al. STREngthening analytical thinking for observational studies: the STRATOS initiative. *Stat Med* 2014;33(30):5413-32.
- 5 Rochon J, Gondan M, Kieser M. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Med Res Methodol* 2012;12(1):81.
- 6 Lotterhos KE, Moore JH, Stapleton AE. Analysis validation has been neglected in the Age of Reproducibility. *PLoS Biol* 2018;16(12):e3000070-e70.
- 7 Brakenhoff TB, Van Smeden M, Visseren FL, et al. Random measurement error: why worry? An example of cardiovascular risk factors. *PLoS One* 2018;13(2):e0192298.
- 8 Abrahamowicz M, Beauchamp M-E, Fournier P, et al. Evidence of subgroup-specific treatment effect in the absence of an overall effect: is there really a contradiction? *Pharmacoepidemiology and Drug Safety* 2013;22(11):1178-88.
- 9 Boulesteix A-L, Wilson R, Hapfelmeier A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. *BMC Med Res Methodol* 2017;17(1):138-38.
- 10 Boulesteix A-L, Binder H, Abrahamowicz M, et al. On the necessity and design of studies comparing statistical methods. *Biometrical Journal* 2018;60(1):216-18.
- 11 Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 1998;17(8):857-72.
- 12 Shmueli G. To Explain or to Predict? *Statist Sci* 2010;25(3):289-310.
- 13 De Bin R, Janitzka S, Sauerbrei W, et al. Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics* 2016;72(1):272-80.

- 1  
2  
3 14 Boulesteix AL, Wright MN, Hoffmann S, et al. Statistical learning approaches in the  
4 genetic epidemiology of complex diseases. *Hum Genet* 2020;139:73-84.  
5 15 De Bin R, Boulesteix AL, Benner A, et al. Combining clinical and molecular data in  
6 regression prediction models: insights from a simulation study. *Brief Bioinform* 2020.  
7 16 Graf E, Schmoor C, Sauerbrei W, et al. Assessment and comparison of prognostic  
8 classification schemes for survival data. *Stat Med* 1999;18(17-18):2529-45.  
9 17 Boulesteix AL, Strobl C, Augustin T, et al. Evaluating microarray-based classifiers: an  
10 overview. *Cancer Inform* 2008;6:77-97.  
11 18 Hanczar B, Hua J, Dougherty ER. Decorrelation of the true and estimated classifier errors  
12 in high-dimensional settings. *EURASIP J Bioinform Syst Biol* 2007:38473.  
13 19 Kent P, Jensen RK, Kongsted A. A comparison of three clustering methods for finding  
14 subgroups in MRI, SMS or clinical data: SPSS TwoStep Cluster analysis, Latent Gold  
15 and SNOB. *BMC Med Res Methodol* 2014;14:113-13.  
16 20 Coretto P, Hennig C. A simulation study to compare robust clustering methods based on  
17 mixtures. *Advances in Data Analysis and Classification* 2010;4(2):111-35.  
18 21 Zipf G, Chiappa M, Porter KS, et al. National health and nutrition examination survey:  
19 plan and operations, 1999-2010. *Vital Health Stat 1* 2013(56):1-37.  
20 22 Brakenhoff TB, Mitroiu M, Keogh RH, et al. Measurement error is often neglected in  
21 medical literature: a systematic review. *Journal of clinical epidemiology* 2018;98:89-  
22 97.  
23 23 Carroll RJ, Ruppert D, Stefanski LA, et al. *Measurement Error in Nonlinear Models: A  
24 Modern Perspective, Second Edition*: CRC Press, 2006.  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8 **FIGURE CAPTIONS**  
9

10 **Figure 1. Estimates of the association between HbA1c levels and systolic blood**  
11 **pressure after adjustment for confounding by BMI under various simulation scenarios**  
12 **characterised by different levels of measurement error. Numbers represent effect**  
13 **estimates averaged over 1,000 simulation runs. Confidence intervals are omitted for clarity.**  
14  
15 *See text for details.*  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



Estimates of the association between HbA1c levels and systolic blood pressure after adjustment for confounding by BMI under various simulation scenarios characterised by different levels of measurement error.

186x126mm (96 x 96 DPI)

```

1
2 # =====
3 # R CODE
4 # small scale simulation study to investigate impact of measurement error
5 # measurement error on (continuous) exposure and/or (continuous) confounding variable
6 # =====
7 #
8 # libraries:
9 library(Hmisc)
10 library(mice)
11 library(tidyverse)
12 #setwd("")
13 # =====
14 # set working directory:
15 # setwd("")
16 # =====
17 # The data can be downloaded in xpt form from https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015
18 nYear=2015
19 # read data:
20 d1 <- sasxport.get("DEMO_I.xpt")
21 d2 <- sasxport.get("BPX_I.xpt")
22 d3 <- sasxport.get("BMX_I.xpt")
23 d4 <- sasxport.get("GHB_I.xpt")
24 d5 <- sasxport.get("TCHOL_I.xpt")
25
26 d1.t <- subset(d1,select=c("seqn", "riagendr", "ridageyr"))
27 d2.t <- subset(d2,select=c("seqn", "bpxsy1"))
28 d3.t <- subset(d3,select=c("seqn", "bmxbmi"))
29 d4.t <- subset(d4,select=c("seqn", "lbggh"))
30 d5.t <- subset(d5,select=c("seqn", "lbdtsi"))
31
32 d <- merge(d1.t,d2.t)
33 d <- merge(d,d3.t)
34 d <- merge(d,d4.t)
35 d <- merge(d,d5.t)
36 # =====
37 # rename variables:
38 # RIAGENDR - Gender
39 # RIDAGEYR - Age in years at screening
40 # BPXSY1 - Systolic: Blood pres (1st rdg) mm Hg
41 # BMXBMI - Body Mass Index (kg/m**2)
42 # LBDTCSI - Total Cholesterol (mmol/L)
43 # LBXGH - Glycohemoglobin (%)
44
45 d$age <- d$ridageyr
46 d$sex <- d$riagendr
47 d$bp <- d$bpxsy1
48 d$bmi <- d$bmxbmi
49 d$HbA1C <- d$lbggh
50 d$chol <- d$lbdtsi
51 d$age[d$age<18] <- NA
52 # =====
53 # select complete cases:
54 dc <- cc(subset(d,select=c("age", "sex", "bmi", "HbA1C", "bp")))
55
56 # analysis:
57 summary(lm(bp ~ HbA1C + age + as.factor(sex), data=dc))
58 confint(lm(bp ~ HbA1C + age + as.factor(sex), data=dc))
59
60 summary(lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc))
61 confint(lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc))

```



```

1
2
3
4 # =====
5 # simulation of measurement error:
6 ref <- lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc)$coef[2]
7 n.sim <- 1e3
8 perc.me.exp <- seq(0,.5,.1)
9 perc.me.conf<- seq(0,.5,.1)
10 scenarios <- expand.grid(perc.me.exp,perc.me.conf)
11 var.exp <- var(dc$HbA1C)
12 var.conf <- var(dc$bmi)
13 n <- dim(dc)[1]
14 beta.hat <- matrix(ncol=dim(scenarios)[1], nrow=n.sim)
15
16 for (k in 1:n.sim){
17   print(k)
18   set.seed(k)
19   for (i in 1:dim(scenarios)[1]){
20     var.me.exp <- var.exp*scenarios[i,1]/(1-scenarios[i,1])
21     var.me.conf <- var.conf*scenarios[i,2]/(1-scenarios[i,2])
22     dc$HbA1C.me <- dc$HbA1C + rnorm(dim(dc)[1], 0, sqrt(var.me.exp) )
23     dc$bmi.me <- dc$bmi + rnorm(dim(dc)[1], 0, sqrt(var.me.conf) )
24     beta.hat[k,i] <- lm(bp ~ HbA1C.me + age + bmi.me + as.factor(sex), data=dc)$coef[2]
25   }
26 }
27 # =====
28 # create figure:
29 tot.mat <- cbind(100*scenarios,apply(beta.hat,2,mean))
30 colnames(tot.mat) <- c("me.exp","me.conf","estimate")
31
32 FIGURE <- ggplot(tot.mat, aes(me.exp, me.conf)) +
33   geom_tile(color="white",aes(fill = estimate)) +
34   geom_text(aes(label = round(estimate, 2))) +
35   scale_fill_gradient2(low="#D55E00",mid="white",high = "#56B4E9", midpoint=ref) +
36   labs(x=paste("% of total variance of HbA1c due to measurement error"),
37        y=paste("% of total variance of BMI due to measurement error")) +
38   coord_equal()+
39   scale_y_continuous(breaks=unique(tot.mat[,1]))+
40   scale_x_continuous(breaks=unique(tot.mat[,1]))+
41   theme(panel.background = element_rect(fill='white', colour='grey'),
42         plot.title=element_text(hjust=0),
43         axis.ticks=element_blank(),
44         axis.title=element_text(size=12),
45         axis.text=element_text(size=10),
46         legend.title=element_text(size=12),
47         legend.text=element_text(size=10))
48
49 FIGURE
50 # savePlot("Figure_STRATOS.tif", type="tif")
51 # =====
52 # END OF R CODE
53 # =====
54
55
56
57
58
59
60

```

# BMJ Open

## An introduction to statistical simulations in health research

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2020-039921.R1
Article Type:	Communication
Date Submitted by the Author:	08-Oct-2020
Complete List of Authors:	Boulesteix, Anne-Laure; Ludwig-Maximilians-Universitat Munchen, Institute for Medical Information Processing, Biometry and Epidemiology Groenwold, Rolf; LUMC Abrahamowicz, Michal; Division of Clinical Epidemiology, McGill University Health Centre Binder, Harald; University of Freiburg, Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center Briel, Matthias; University Hospital Basel, Institute for Clinical Epidemiology and Biostatistics Hornung, Roman; Ludwig-Maximilians-Universitat Munchen, Institute for Medical Information Processing, Biometry and Epidemiology Morris, Tim; MRC Clinical Trials Unit at UCL, ; Rahnenführer, Jörg; TU Dortmund, Department of Statistics Sauerbrei, Willi; University of Freiburg Hospital, Institute for Medical Biometry and Statistics
<b>Primary Subject Heading</b>:	Epidemiology
Secondary Subject Heading:	Research methods
Keywords:	STATISTICS & RESEARCH METHODS, EPIDEMIOLOGY, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT

SCHOLARONE™  
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

## An introduction to statistical simulations in health research

Anne-Laure Boulesteix<sup>1</sup>, Rolf Groenwold<sup>2,3</sup>, Michal Abrahamowicz<sup>4</sup>, Harald Binder<sup>5</sup>, Matthias Briel<sup>6,7</sup>, Roman Hornung<sup>1</sup>, Tim Morris<sup>8</sup>, Jörg Rahnenführer<sup>9</sup>, Willi Sauerbrei<sup>5</sup>

on behalf of the Simulation Panel of the STRATOS initiative

\*To whom correspondence should be addressed: Anne-Laure Boulesteix, IBE, Marchioninstr. 15, 81377 Munich, Germany. [boulesteix@ibe.med.uni-muenchen.de](mailto:boulesteix@ibe.med.uni-muenchen.de), +49 89 440077598

<sup>1</sup> Institute for Medical Processing, Biometry and Epidemiology, Ludwig Maximilian University of Munich, Munich, Germany

<sup>2</sup> Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, the Netherlands

<sup>3</sup> Department of Biomedical Data Science, Leiden University Medical Centre, Leiden, the Netherlands

<sup>4</sup> Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada

<sup>5</sup> Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany

<sup>6</sup> Department of Clinical Research, Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel and University of Basel, Basel, Switzerland

<sup>7</sup> Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada

<sup>8</sup> MRC Clinical Trials Unit at UCL, London, UK

<sup>9</sup> Department of Statistics, TU Dortmund University, Dortmund, Germany

Word count: 6556

**ABSTRACT**

In health research, statistical methods are frequently used to address a wide variety of research questions. For almost every analytical challenge, different methods are available. But how do we choose between different methods and how do we judge whether the chosen method is appropriate for our specific study? Like in any science, in statistics, experiments can be run to find out which methods should be used under which circumstances. The main objective of this paper is to demonstrate that simulation studies, i.e., experiments investigating synthetic data with known properties, are an invaluable tool for addressing these questions. We aim to provide a first introduction to simulation studies for data analysts or, more generally, for researchers involved at different levels in the analyses of health data, who (i) may rely on simulation studies published in statistical literature to choose their statistical methods and who, thus, need to understand the criteria of assessing the validity and relevance of simulation results and their interpretation; and/or (ii) need to understand the basic principles of designing statistical simulations in order to efficiently collaborate with more experienced colleagues or start learning to conduct their own simulations. We illustrate the implementation of a simulation study and the interpretation of its results through a simple example inspired by recent literature, which is completely reproducible using the R-script available from the supplement.

## 1 INTRODUCTION

In health research, statistical methods are frequently used to address a wide variety of research questions. For almost every analytical challenge, different methods are available. But how do we choose between different methods and how do we judge whether the chosen method is appropriate for our specific study? Most statistical methods are developed under specific assumptions, but these assumptions are often difficult to check in applied settings. Moreover, performance of methods may still be reasonable when some assumptions are violated, such as the linearity of relationships in regression models in the presence of mild non-linear relationships. In real-life studies of human health, some of these formal underlying assumptions may be questionable or definitely violated. For example, frequent problems, such as unusual distributions, missing data, measurement errors, unmeasured confounders, or lack of accurate information on event times, may affect the accuracy or even the validity of the proposed analyses. What conditions (e.g., what sample size) are needed for a specific method to behave well? Which method is most appropriate in a particular setting?

The main objective of this paper is to demonstrate that simulation studies, i.e., evaluation of synthetic data with known properties, are an invaluable tool for addressing these questions. We aim to provide a first introduction to simulation studies for data analysts or, more generally, for researchers involved at different levels in the analyses of health data, for example, data from observational studies or from clinical trials, who (i) may rely on simulation studies published in statistical literature to choose their statistical methods and who, thus, need to understand the criteria of assessing the validity and relevance of simulation results and their interpretation; and/or (ii) need to understand the basic principles of designing statistical simulations in order to efficiently collaborate with a more experienced colleague or to start learning to conduct their own simulations. Our paper is intended for an audience that is otherwise not targeted by previous literature on simulation studies and uses a novel approach to introduce the basic principles of simulation studies to clinical researchers and end users of statistical methods. Statisticians interested in more details about statistical simulations are referred to the more technical overviews available in the literature.[1-3]

More generally, our introduction to simulation studies aims to draw the attention of readers of medical papers, including practitioners, to the importance of the choice of appropriate, validated statistical methods. The use of inappropriate statistical methods contributes to the replication crisis that has drawn increasing attention in recent years; see for example the Lancet series "Increasing value, reducing waste".[4] Simulation studies have a role to play in this global process as they are a means of identifying the appropriate methodology for a particular study in a specific context, thus improving research quality. In this context, understanding the principles of simulation studies allows clinical researchers to better use published simulation results. Note that simulation studies themselves also have to be relevant and replicable.

1  
2  
3 Statistical methodology has seen substantial development in recent times but many of these  
4 developments are largely ignored in the practice of health data analyses. To help bridge the  
5 gap between methodological innovation and applications to medical data, the  
6 STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative was  
7 launched in 2013.[5] It aims to provide statistical guidance for key topics in the design and  
8 analysis of observational studies. In practice, analyses are sometimes conducted by  
9 researchers with limited statistical background. Consequently, STRATOS plans to develop  
10 guidance for researchers with different levels of statistical knowledge including researchers  
11 without strong statistical backgrounds (see Table 1 in [5]). For the analysis of observational  
12 studies, typically several approaches are possible, and the properties of each approach  
13 should be assessed in comparison with alternative methods. Simulation studies are key  
14 instruments for such assessments. Ideally, all data analysts should be familiar with them.

15  
16  
17 This paper is structured as follows. We first discuss the role of statistical simulation studies in  
18 section 2. Section 3 outlines four relatively simple examples of statistical methods and then  
19 explains how the performance of these methods could be evaluated using simulation  
20 studies. Section 4 sketches out the basic principles of designing and conducting simulations.  
21 Finally, section 5 briefly illustrates the implementation of a simulation study and the  
22 interpretation of its results through a simple example inspired by recent literature.

## 2 THE ROLE OF SIMULATION STUDIES

### Comparing methods based on theory

23  
24  
25 During the first half of the 20th century, mathematical theory was the cornerstone of  
26 evaluating traditional statistical methods addressing well defined problems. However, to  
27 investigate questions in modern medicine, more complex statistical modelling or the use of  
28 machine learning techniques are often required. Only in rare cases of low complexity and  
29 often of limited practical relevance, mathematics tells us that - given the data satisfy certain  
30 properties - the considered method behaves in a particular way. For example, theory tells us  
31 that the two-sample t-test has better power to detect a true difference between mean values  
32 in two independent groups than the Mann-Whitney test - if the variable of interest is normally  
33 distributed within each of the two groups. Most theoretical results of this type are valid only  
34 under specific assumptions about the available data. While it may be acceptable to assume  
35 normally distributed data in the case of the simple example mentioned above, for more  
36 complex problems the required assumptions can be unrealistic; see sections 3.2, 3.3 and 3.4  
37 for examples beyond this simple case. Moreover, the process of verifying assumptions is  
38 often already challenging in practice; see for example [6] for an extensive simulation study of  
39 the choice between t-test and Mann-Whitney test including considerations on normality  
40 checks.

## Comparing methods using empirical data

Another approach for evaluating statistical methods consists of applying them to representative datasets from the considered field and assessing their performance; or, more generally, of observing their behaviour when using them in these datasets. Some important characteristics of statistical methods can indeed be derived from real datasets. For example, are results stable if we modify the dataset slightly? For many approaches, however, the most important evaluation criteria cannot be assessed for real data, simply because for real data we do not know the true values of the underlying parameters we aim to draw inferences about. For example, if one method estimates a difference of 1 between two groups, and another estimates a difference of 2, we can see that they give us different results (assuming that the confidence intervals are narrow), but we do not know whether 1 or 2 is closer to the correct answer.

## Why simulation studies?

A simulation study is useful if theoretical arguments are insufficient to determine whether the method of interest is valid in a specific real-life application or whether violations of the assumptions underlying the available theory (such as normal distribution of residuals, proportional hazards, etc.) affect the validity of the results. In methodological research, simulations play a role similar to experiments in basic science.[7] The idea of a simulation study is to investigate the behaviour of methods when applied to synthetic datasets with known characteristics. Because the 'correct' or 'true' answer is known by the researchers, who had full control of how the data were simulated, simulations permit assessment of whether the methods recover this known truth. For example, we may generate data with and without a treatment effect and then assess how often a test correctly or incorrectly rejects the null hypothesis of no treatment effect. Alternatively, we may generate data in which the treatment effect has a certain value and then study how accurately a regression model can estimate this known effect. Notice that such assessment is *not* possible using real data when the true response or the true effect is not known.

Suppose a scientist is planning a cohort study of the effect of an exposure on time to a clinical event (e.g., death) and wants to know what sample size is necessary to achieve a certain power with a given test, or a certain precision with a given estimation method. A question that might be explored using a simulation study could be the following: What is the power of the logrank-test (an asymptotic test requiring large sample sizes to ensure validity), in the case of small samples? Here, a simple simulation study, designed to be consistent with the specific settings of the proposed study (sample size, prevalence of the exposure of interest, incidence of events, etc.), could provide the necessary answers.

Simulation studies are also helpful to provide objective reproducible answers to more general methodological questions on the behaviour of statistical methods (i.e., not necessarily motivated through a specific application). Examples of this type of question, which have been investigated by recent simulation studies, include: What is the effect of



1  
2  
3 measurement errors on the estimated exposure-outcome relations in epidemiological  
4 studies?[8] Does it make sense to check for subgroup-specific treatment effects even if the  
5 test for an overall effect is non-significant?[9]  
6  
7

8 In addition to the evaluation of individual methods, simulations can also be used to  
9 determine which one of several candidate methods will perform best for the application at  
10 hand. In the case of simulations reported in statistical literature, candidate methods may  
11 include existing methods, and may (but do not have to) include new methods proposed by  
12 the researchers performing the simulation study. In the latter case, their focus is often on  
13 showing in which settings the new method performs better than its existing ‘competitors’.[10,  
14 11]  
15  
16  
17

18 No matter the context of the simulation study, the objective is to find out if/when methods  
19 perform well and when they fail. Regarding the “when” question, simulations provide an ideal  
20 setting for a systematic assessment of how variations in the values of relevant parameters,  
21 and/or assumptions regarding data structure (e.g., independence of observations, lack of  
22 measurement errors) affect the performance of the methods of interest. The definition of the  
23 term “good performance” depends on the context. For example, if we compute a 95%  
24 confidence interval, we usually want it to yield 95% coverage (i.e., we want 95% of the  
25 confidence intervals constructed in this way, using varying datasets, to cover the true value).  
26 If we apply a statistical test, we want this test to reject the null hypothesis with high  
27 probability if it is false, but to *retain* it with high probability if it is true. In comparison studies,  
28 two or more methods may be compared in this respect. In the case of a simulation  
29 performed for sample size calculation, we want to determine the smallest sample size with  
30 which a study has a given power to detect clinically important effects.  
31  
32  
33  
34  
35  
36

37 In practice, nobody can predict with certainty whether a method will yield accurate results for  
38 a specific dataset, or which of a set of considered methods will perform best on that dataset.  
39 Simulations can provide *systematic evidence* regarding how methods perform on average for  
40 datasets with similar characteristics to the dataset under investigation. In an ideal world,  
41 relevant results from simulation studies would be available from previous research to help  
42 make rational decisions about which method to use. Data analysts would then use simulation  
43 results to verify whether the method they choose is adequate, or to pick the most suitable  
44 from a range of different methods. Such “previous research” is typically done by statistical  
45 researchers working on methods as the focus of research (as opposed to researchers  
46 *applying* methods in health research projects). For a data analyst with little experience and  
47 background in statistical methodological research, it is important to be able to interpret the  
48 results of such simulation studies. If previous evidence is lacking, or if previous studies do  
49 not seem to apply to the specific data setting under consideration, data analysts should  
50 conduct a targeted simulation study tailored to their specific dataset.  
51  
52  
53  
54  
55  
56  
57

### 58 **3 EXAMPLES OF STATISTICAL METHODS**

59  
60

1  
2  
3 In this section we present four examples of analyses which help us to explain the basic  
4 principles of simulation studies. Key criteria for evaluating the performance of methods  
5 related to these examples are summarised in Table 1, at the end of the section.  
6  
7

### 8 **3.1 Statistical hypothesis testing and confidence intervals**

9

10 In most health research projects we perform statistical tests and/or derive confidence  
11 intervals. However, their behaviour is often not well-characterised in real world situations.  
12 For example, for time-to-event data with censored observations, how do the logrank-test and  
13 confidence intervals for the hazard ratio behave in relatively small samples? Which  
14 technique should be preferred to compute confidence intervals for proportions in a given  
15 setting (e.g., very small proportions)?[12]  
16  
17  
18

- 19 ● What is a good test/confidence interval?

20  
21  
22 A good test is one that yields the correct answer with high probability, i.e., one that rejects  
23 the null hypothesis with high probability if it is not true, and retains it with high probability if it  
24 is true. Classical tests are defined in such a way that, in theory, the probability that the null  
25 hypothesis is rejected despite being true (called type 1 error) does not exceed a level  $\alpha$   
26 chosen by the user (in medicine, often  $\alpha=0.05$ ) - provided the assumptions are fulfilled.  
27 However, it is possible that the actual type 1 error may be larger than  $\alpha$ , in which case the  
28 results of the test should be interpreted with caution. When evaluating a test, it is thus  
29 important to verify that the type 1 error does not exceed the nominal significance level  $\alpha$  that  
30 was chosen by the researcher. Provided the type 1 error is as it should be (equal to or  
31 smaller than  $\alpha$ ), the most important quantity characterising a statistical test is its power,  
32 defined as the probability of correctly rejecting the null hypothesis.  
33  
34  
35  
36  
37

38 Apart from hypothesis testing, results of statistical analysis are often presented as an  
39 estimate with a corresponding confidence interval. A good method for deriving, say, 95%  
40 confidence intervals is a method that yields confidence intervals covering the true value with  
41 probability 95%.  
42  
43  
44

- 45 ● Can real data be used for the evaluation?

46  
47 The main performance criteria cannot simply be assessed based on real data, because the  
48 truth (which hypotheses are true or false, or the true value of the parameter being estimated)  
49 is generally unknown in practice - we can see that a test has rejected the null hypothesis, but  
50 do not know if this was correct or not. If the truth were known, there would be no need to  
51 perform the test or compute a confidence interval. Baseline characteristics in correctly  
52 randomised trials are a notable exception. Given the randomisation procedure, they are  
53 expected to be equally distributed in the two groups by definition.  
54  
55  
56  
57

### 58 **3.2 Model selection for regression models: explaining the effects of covariates on an** 59 **outcome variable**

60

1  
2  
3 The second example is regression modelling of an outcome variable of interest, sometimes  
4 called “dependent” variable, using several covariates, sometimes denoted as predictor  
5 variables or independent variables (often, prognostic or risk factors). In general, such  
6 modelling is performed either to *explain* the outcome variable by determining the effects of  
7 the covariates (as considered in this section), or to build a model, which will be used later on  
8 new patients for *prediction* purposes (as considered in the next section); see [13] for a  
9 discussion of these two related but distinct purposes. In health research, the outcome  
10 variable is often of one of the three following types: continuous (e.g., amount of cholesterol  
11 reduction), categorical (e.g., response to therapy) or survival time (e.g., disease free survival  
12 in months). Even though for all three cases standard regression modelling is reasonably  
13 well-understood, the behaviour of regression techniques (including model selection) still  
14 raises questions in particular cases; see for example a recent simulation study on the use of  
15 resampling techniques for model selection purposes.[14]

- 21 • What is a good regression approach?

22  
23 In principle, a regression technique (including model selection aspects) is expected to (i)  
24 correctly distinguish the variables that are related to the outcome variable from those that are  
25 not, and (ii) correctly fit the regression coefficients of the variables, i.e., fit them to provide  
26 estimated values close to the true ones (unbiased and low variance). Regarding (i), it is good  
27 to have high sensitivity (i.e., selecting most/all variables with effects, this is analogous to  
28 detecting most/all diseased patients in a diagnostic study) as well as high specificity (i.e., not  
29 selecting variables without an effect, analogous to correctly identifying participants without  
30 disease). Depending on the specific goal, analysts may also aim to eliminate variables with  
31 very small effects.

- 32 • Can real data be used for the evaluation?

33  
34 In practice, the exact set of variables that have an effect on the outcome variable and the  
35 values of these effects are unknown, although previous knowledge from the literature may  
36 provide valuable guidance in some cases. Thus, in most cases, real data are of limited use  
37 for the evaluation of model selection approaches for regression models.

### 38 39 40 **3.3 Model selection for regression models: predicting the values of an outcome using 41 the values of covariates**

42  
43 The third example is related to the second example, but takes a different perspective. While  
44 regression models are often used to “explain” the outcome variable (e.g., a disease outcome  
45 or survival time), in order to understand how different risk factors affect the outcome variable,  
46 they can also be used as “prediction models” to predict the outcome of interest for new  
47 patients, based on these patients’ values of the covariates. Classical linear regression  
48 models can be used for this purpose as well as various more complex alternative  
49 procedures, especially algorithms developed in the machine learning community, such as  
50 support vector machines or random forests (see [15] for a gentle introduction). In this field,  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 simulations can be useful to assess the prediction accuracy of the considered prediction  
4 methods in different settings. For example, different penalised regression methods may be  
5 compared in simulations with respect to their prediction performance when a small number  
6 of clinical covariates are combined with a large number of candidate molecular  
7 covariates.[16]  
8  
9

- 10  
11 • What is a good prediction model?

12  
13 A good prediction model is a model that yields accurate predictions in the future patients it  
14 will be applied to. For continuous and for categorical outcome variables, often predicted and  
15 true values are directly compared, and the differences are summarised across patients. For  
16 survival times, suitable adjusted scores, like the Brier score, may be used to take into  
17 account censoring.[17]  
18  
19

- 20  
21 • Can real data be used for the evaluation?

22  
23 The prediction error can be estimated based on the available dataset using a large (possibly  
24 external) validation dataset if available, or so-called resampling techniques such as cross-  
25 validation.[18] Note that this estimation may be unreliable depending on the context (for  
26 example, the smaller the sample size, the more unstable the cross-validation estimates).[19]  
27 What these evaluations tell us about the methods' accuracy is relevant to the considered  
28 specific real data example(s) but may not be relevant to other settings.  
29  
30  
31

### 32 33 **3.4 Clustering**

34  
35 The last example considered in this paper is clustering, also called cluster analysis. The  
36 objective of clustering is to identify clusters, i.e., "groups" of patients that behave similarly.  
37 For example, clustering methods may be used with the goal of identifying clinically  
38 meaningful subgroups of patients, using magnetic resonance imaging data and clinical data,  
39 among others.[20] Clusters should be constructed in such a way that the values of patients  
40 within a cluster are more similar (according to the chosen similarity criterion) than values of  
41 patients from different clusters. Many different clustering algorithms have been proposed at  
42 the interface between computer science and statistics, for example k-means clustering or  
43 hierarchical clustering. Simulation studies may be used to assess the ability of methods to  
44 recover a true underlying structure.[20, 21]  
45  
46  
47  
48

- 49  
50 • What is a good clustering method?

51  
52 A good clustering procedure is a procedure that correctly recovers a true cluster structure  
53 present in the data but does not falsely identify clusters that are not in fact present.  
54  
55

- 56  
57 • Can real data be used for the evaluation?

58  
59 In practice, the true cluster structure is often unknown. And even if there is a known cluster  
60 structure, further sensible cluster structures might exist. The abilities of clustering methods to

group similar observations together may be assessed by using data that consists of known subgroups and measuring the degree of overlap between the clustering structure defined by the known subgroups and the clustering structure proposed by the clustering algorithm. However, there might not be only one sensible cluster structure; in fact, the observations may cluster together more strongly according to factors other than the subgroup membership, e.g., gene expressions are associated with various phenotypes. Real data may be used to assess aspects such as stability (i.e., robustness against small changes in the data) or computational efficiency, but they are of limited use for the evaluation of a clustering method according to the criterion “agreement with the true cluster structure”.

Example	Evaluation criterion	Target value
<b>A – testing and confidence intervals</b>	type 1 error type 2 error coverage of (1- $\alpha$ ) confidence intervals	close to and not greater than nominal value $\alpha$ low close to and not lower than nominal value 1- $\alpha$
<b>B – explaining</b>	mean coefficient values precision of coefficient estimation coverage of confidence interval sensitivity of variable selection specificity of variable selection	close to true values (low bias) high (low variance) close to and not lower than nominal value 1- $\alpha$ high high
<b>C – predicting</b>	prediction error on independent data accuracy measures	low high
<b>D – clustering</b>	agreement with true cluster structure	high
<b>All settings</b>	stability	high

	computational cost	low
	success of the computation (e.g., “convergence”)	yes

Table 1. Overview of the main criteria for evaluating statistical methods in the four considered examples. The last column indicates which values the considered evaluation criterion takes if the investigated method is good.

## 4 BASIC PRINCIPLES OF SIMULATION STUDIES

### 4.1 Key features of a simulation study

In this section we give a brief overview of the key features of a simulation study, which are also displayed in Table 2 together with the example from section 5. A more detailed introduction to the concepts of data generating mechanisms and simulation scenarios is given in section 4.2, for interested readers. One may also refer to a recent in-depth article on simulation studies addressing an audience of statisticians.[3]

The first key feature of a simulation study is its *overall objective*. Is the simulation study tailored to a specific dataset relevant to a particular application or does it address a methodological question of general interest for future applications? Regardless of the overall objective, researchers performing a simulation study should make decisions considering the following key issues.

#### *Aims:*

What do we want to learn about the method(s) from the simulation study? For example, one may want to assess whether a model selection method selects the right covariates (main aim), and whether it estimates their effects accurately (secondary aim). This point is analogous to the definition of primary and secondary outcomes in clinical trials, e.g., disease-free survival or side effects.

#### *Data generating mechanism (including choice of relevant parameters):*

How do we generate the simulated datasets? From which distribution? Which parameters may affect the results and what values should be considered? Each combination of the relevant assumptions and parameter values defines one simulation scenario (for which several datasets will usually be (randomly) generated, as outlined in the next section). There are many ways to generate datasets: by using real datasets as a starting point (see section 5 for an example) or by sampling from (possibly multivariate) pre-specified distributions, e.g., the normal distribution. The definition of the scenarios is analogous to the definition of

1  
2  
3 experimental conditions for a lab experiment, and should be guided by considerations about  
4 clinical plausibility and/or relevance.[11] While simulation designs can be made complex, the  
5 focus is often on relatively simple properties of the data distributions, such as skewness or  
6 outliers. The performance of many widely used basic statistical building blocks, such as the  
7 least squares optimisation principle for estimating model parameters, can be severely  
8 affected by the type of distribution under consideration. As a result, in order to  
9 comprehensively gauge performance, simulation studies should also include the rather  
10 innocent looking problems of real data, such as some outlier observations. More insights are  
11 given in section 4.2.

12  
13  
14  
15  
16 *Method(s) of analysis to be evaluated/compared:*

17  
18 Which method(s)/variant(s) is (are) evaluated? This point is analogous to the definition of the  
19 treatments with all necessary details (dose, etc.) to be compared in a clinical trial. Further  
20 discussion about the analogy between clinical trials and comparisons of statistical methods  
21 can be found elsewhere.[10]

22  
23  
24  
25 *Performance measure(s):*

26  
27 Which criteria are used to assess the performance of the considered data analysis methods?  
28 In the example of model selection mentioned above, one may address the main aim by  
29 considering the sensitivity of the method for selecting the “true effects” as well as the  
30 frequency of “false positives” (i.e., selection of variables that have no true associations with  
31 the outcome). The secondary aim may be addressed by computing the mean squared  
32 deviation or the mean absolute deviation of the coefficient estimates from the true values.  
33 This point is analogous to the precise definition of primary and secondary outcomes in a  
34 clinical trial: e.g., which instruments are used for the assessment of side effects of the  
35 therapy, or how do we exactly estimate disease-free survival and compare it across the trial  
36 arms?

37  
38  
39  
40  
41  
42 *Number of repetitions:*

43  
44 For each considered scenario, how many datasets are randomly drawn? It is necessary to  
45 generate several (ideally, “many”) datasets in order to average out random fluctuations and  
46 ensure sufficiently precise simulation results. The more datasets are generated, the more  
47 precise the performance evaluation will be - as can be quantified through, for example, the  
48 width of the confidence intervals for the selected “performance criteria”. The number of  
49 repetitions is analogous to the sample size in a clinical trial. In contrast to increasing the  
50 sample size in clinical trials, however, it is often easy to extend the number of repetitions in  
51 simulation studies. The number of repetitions is chosen as a compromise between precision  
52 of the results and computational time.

Key features of simulation studies	NHANES example
Aims	To quantify the impact of measurement error
Data-generating mechanism	Take real data, add normally-distributed random error to the exposure of interest (HbA1c) and/or the confounder (BMI)
Method of analysis	Multivariable linear regression, first on data with no measurement error, then on data with measurement error added
Performance measure	Bias in regression coefficient for exposure of interest (HbA1c)
Number of repetitions	1,000

Table 2. Overview of the key features of a simulation study (1<sup>st</sup> column) with the NHANES example described in section 5 (2<sup>nd</sup> column). This table is inspired from the “ADEMP” system (Aims, Data-generating mechanisms, Estimands, Methods, and Performance measures) introduced previously in statistical literature.[3]

#### 4.2 Sampling variability and data generating processes

This section gives further insights into the data generating process for readers interested in gaining a deeper understanding of the fundamentals of simulation studies, beyond the key points outlined above. To this end we first explain briefly how simulations provide a framework for assessing and accounting for the impact of random sampling error on the results of empirical studies.

##### *Preliminary: Sampling variability in real data:*

Suppose a clinical researcher is interested in the mean difference between the blood pressure of males and females in the population aged 20 to 60. The true mean difference could only be calculated if we had data on the whole populations of males and females aged 20 to 60. Of course, in practice, we only have a sample available with a specific (often moderate) size and can only *estimate* the mean difference using this sample. Different samples will yield different estimates of the same mean difference in the population. Collecting a data sample can be seen as drawing observations from a population of interest that has particular characteristics. In statistical terms, these observations can be seen as random observations generated from the *true distribution of the variable(s) of interest* in the



1  
2  
3 relevant population. In real-life studies, this distribution and the true values of its parameters  
4 (e.g., population means) are unknown and we can only *estimate* them using available  
5 sample data.  
6  
7

#### 8 *Simulating data:*

9  
10 The principle of simulations is to mimic the process of taking repeated (random) samples  
11 from a large population, by repeatedly generating synthetic data (“virtual observations”) from  
12 a virtual population, under pre-specified assumptions that can be varied across the  
13 considered simulation scenarios. Each synthetic sample is generated from a particular  
14 known distribution, with “true” values of all relevant parameters fixed by the researchers.  
15 Each simulated sample is then analysed using the method(s) of interest, and its (their)  
16 performance is evaluated using pre-specified criteria (see Table 1 for examples). To give  
17 one simple example, we may simulate systolic blood pressure (SBP) values for a sample of  
18 n=100 “synthetic subjects” by generating 100 independent numbers from a normal  
19 distribution with, say, mean 120 and standard deviation 15. Doing so, we know that the true  
20 population mean is 120 mm Hg and that the simulated blood pressure follows the normal  
21 distribution. The way in which virtual observations are generated in the context of a  
22 simulation (in our example, “100 independent numbers from a normal distribution with mean  
23 120 and standard deviation 15”) is termed the *data generating mechanism*. There is a large  
24 number of user-friendly statistical packages that can be used to accomplish this task.  
25  
26  
27  
28  
29  
30  
31

#### 32 *Sampling variability in simulations:*

33  
34 Just as random sample-to-sample variability affects real data samples drawn from a  
35 population of interest, it also affects the results obtained using simulated data. If we generate  
36 two synthetic datasets using the same data generating mechanism and the same  
37 parameters, we will get somewhat different results (with the differences decreasing, on  
38 average, with increasing size of the generated datasets). It is therefore almost always  
39 important to repeat the same data generation and analysis process using many simulated  
40 datasets, as outlined in the paragraph “*Number of repetitions*” of section 4.1 above. The  
41 variability of the results obtained across the different datasets simulated from the same  
42 distribution has to be carefully assessed by, for example, calculating the standard deviation  
43 of the individual estimates. Calculating the mean value of the individual estimates provides a  
44 more robust estimate of the unknown population-level parameter than a value from a single  
45 simulated sample, as averaging over several repetitions reduces the impact of random  
46 sampling error.  
47  
48  
49  
50  
51  
52

#### 53 *Choice of data generating mechanisms:*

54  
55 When performing a simulation, one has to choose one or several data generating  
56 mechanisms that reflect, as closely as possible, the distribution and relevant characteristics  
57 of the real data of interest, no matter whether the focus is on a specific application or on a  
58 ‘generic’ methodological question such as evaluation or comparison of specific analytical  
59  
60

1  
2  
3 methods. The difficulty is that, in reality, the true data generating process is unknown as  
4 mentioned above in the example of blood pressure. The only possibility is to consider  
5 several data generating mechanisms – called simulation scenarios – that, together, will cover  
6 the range of situations congruent with the expected structure of real data of interest.  
7 Scenarios may differ, among other ways, in the sample size, the true distributions of the  
8 considered variables (normal, uniform, exponential, etc.), the values of parameters such as  
9 means or variances, the correlation structure of the variables or the presence of outliers. For  
10 example, we may be interested in the behaviour of a test that assumes a normal distribution  
11 in situations where this assumption is not fulfilled. If the variable of interest is expected,  
12 based on earlier studies and/or substantive knowledge, to be (approximately) uniformly  
13 distributed (meaning that the observations are evenly distributed over a certain interval),  
14 priority will be given to corresponding scenarios. However, it may be useful to also consider  
15 a few alternative scenarios with other distributions, e.g., a positively skewed distribution with  
16 most values concentrating below the mean and relatively few high values.  
17  
18  
19  
20  
21  
22

23 In general, if the focus of the simulation study is on a specific application, the primary goal is  
24 essentially to simulate datasets that are as similar as possible to the relevant real dataset.  
25 This may necessitate making some plausible assumptions and involve some uncertainty if  
26 the data have not yet been collected – as is the case when simulations are performed with  
27 the aim of calculating the adequate sample size or assessing the expected power and/or  
28 precision of future analyses. In contrast, if the focus of the simulation is on the general  
29 behaviour of a particular method (or comparison of alternative methods) for a class of  
30 applications, the primary goal when choosing scenarios is often to cover a wide spectrum of  
31 potentially plausible situations in which the method(s) of interest are likely to be employed.  
32 Some scenarios may be unrealistic but are nevertheless helpful in understanding how the  
33 method works or when it breaks down (and how it can be improved to cope better with the  
34 problematic situations), and thus yield valuable information. The choice of simulation  
35 scenarios is thus intrinsically related to the goal of the simulation, but should also account for  
36 substantive knowledge in the field of potential real-life applications.  
37  
38  
39  
40  
41  
42

### 43 **4.3 Advantages and drawbacks of simulation studies**

44  
45  
46 To simulate the synthetic datasets, we define the underlying “truth” regarding the research  
47 question being explored. For example, in example A in section 3 (testing) we know whether  
48 the null hypothesis is true or not. In example B (explaining) we know which variables have  
49 independent effects on the outcome variable. In example C (predicting) we know the true  
50 values of the outcome variable. In example D (clustering) we know the true cluster structure.  
51 To sum up, in all these examples, we know what an *accurate* method of data analysis is  
52 supposed to find. Thus, we can determine how well the method(s) being evaluated  
53 perform(s) by comparing their results against this known “truth”. This feature is the major  
54 advantage of simulations over empirical comparisons of the same methods based on one or  
55 few real-life datasets as, in the latter case, the true answers often remain unknown.  
56  
57  
58  
59  
60

1  
2  
3 Another advantage of simulations is that they allow investigation of a large number of  
4 different scenarios, and in particular also scenarios that are not directly observed in real  
5 datasets. This means that the analysis can be extended to new or rare scenarios, or  
6 scenarios reflecting practically unrealistic settings (e.g., randomised trial data or very large  
7 sample sizes). A related advantage of simulations is that, by varying the assumptions and  
8 the values of relevant parameters used to generate data for different scenarios, one can  
9 *systematically* assess how the performance of different methods depends on these  
10 assumptions and parameters. Furthermore, one can also perform, for each considered  
11 scenario, as many repetitions as needed to average out random fluctuations. This is in  
12 contrast to real data experiments where the quantity of data is often severely limited, which  
13 affects the precision of the results.  
14  
15  
16  
17  
18

19 These advantages, however, come at a cost. Firstly, simulation scenarios are often  
20 simplified, i.e., do not reflect the true complexity of the data encountered in real-life data  
21 analyses. The lack of complexity of simulated data may lead to a distorted picture of the  
22 methods' performance. For example, an approach that can model data in a very flexible  
23 manner might be more severely affected by outliers. Yet, simulation designs so far rarely  
24 incorporate outliers or skewed distributions. Real-world performance of an approach that has  
25 been selected based on simulation study results might be surprisingly bad. Secondly, large  
26 simulation studies can be computationally very expensive, taking days or weeks and even  
27 requiring the use of parallel computing, if a large number of scenarios and/or large numbers  
28 of repetitions are considered and especially if the analysis also involves large datasets  
29 and/or complex statistical methods.  
30  
31  
32  
33  
34

35 Finally, it is important to note that simulations are not immune to the typical flaws of  
36 numerical studies leading to biased results. For example, the effect of single influential  
37 points, which are difficult to detect in simulation studies with hundreds or thousands  
38 of simulated samples, can be critical. They may be relevant in some of the simulation  
39 repetitions, in which they cause unreliable results. If undetected, they can bias the results.  
40 Most importantly, selective reporting may be an issue. If a very large number of scenarios is  
41 analysed, but only those scenarios that favour one particular method are presented in the  
42 paper, the reported results will give a distorted picture of reality. Obviously, this is a serious  
43 problem of bad reporting and bad research, which can be easily avoided by being  
44 transparent.  
45  
46  
47  
48  
49  
50  
51

## 52 **5 AN EXAMPLE OF A STATISTICAL SIMULATION**

53

54 For illustration, in this section we consider a simple simulation study that investigates the  
55 impact of measurement error in linear regression analysis, inspired by a previous study.[8]  
56 See the overview of its key features in the right column of Table 2. Our study is completely  
57 reproducible using the R code provided in supplementary file 1, which uses freely available  
58 data. In epidemiological studies of the relation between an exposure and an outcome, this  
59  
60

1  
2  
3 relation is often estimated using regression analysis. As an example, we consider a study of  
4 the association between glycated haemoglobin levels (HbA1c) and systolic blood pressure  
5 assessed using linear regression. Data from 5092 subjects in the 2015–2016 National Health  
6 and Nutrition Examination Survey (NHANES)[22] is used to obtain an estimate of the effect  
7 of HbA1C on systolic blood pressure, while adjusting for age, gender, and body mass index  
8 (BMI). Details on the data are described on the NHANES website:  
9 [\[https://wwwn.cdc.gov/nchs/nhanes/\]](https://wwwn.cdc.gov/nchs/nhanes/). After adjustment for age and gender, it was estimated  
10 that HbA1c increases systolic blood pressure by 1.13 mmHg (95%CI 0.73 – 1.52) per unit  
11 increase in HbA1c. Additional adjustment for BMI resulted in a considerable change in the  
12 effect estimate: HbA1c was estimated to increase blood pressure by 0.75 mmHg (95%CI  
13 0.35 – 1.16) per unit increase in HbA1c.  
14  
15  
16  
17  
18

19 The confounding variable BMI as well as the exposure variable HbA1c may be subject to  
20 measurement error. For example, BMI may be self-reported (instead of a standardised  
21 measurement using scales) or technical problems in the lab may have affected the HbA1c  
22 measurement. Therefore, researchers may want to know the possible impact of  
23 measurement error of the exposure and/or confounding variable(s) in terms of bias.[23] We  
24 are interested both in the direction and magnitude of this bias.  
25  
26  
27

28 One way to investigate the possible impact of measurement error is through a small  
29 simulation study[8], whose steps are schematically represented in Figure 1. For the purpose  
30 of this example, the original recordings in the NHANES data were assumed to be measured  
31 without error (step 1 in Figure 1). Then, in addition, new artificial variables were created that  
32 represented HbA1c and BMI, but for the situation in which these are measured with error. To  
33 create these variables, measurement error was artificially added to the exposure variable  
34 (HbA1c) and/or the confounding variable (BMI) (step 2 in Figure 1). These errors were drawn  
35 from a normal distribution with mean zero, and were independent of all variables considered.  
36 This type of measurement error is often referred to as classical measurement error.[24] The  
37 variance of the normal distribution, defining the amount of measurement error added, was  
38 altered for different scenarios. Scenarios ranged from no measurement error on either  
39 HbA1c or BMI (reference scenario) to 50% of the variance in HbA1c and/or BMI attributable  
40 to measurement error. To minimise the impact of simulation error, each scenario was  
41 repeated 1,000 times and results were averaged per scenario over these 1,000 repetitions.  
42  
43  
44  
45  
46  
47

48 Figure 2 shows the impact of measurement error on HbA1c and/or BMI on the estimate of  
49 the regression coefficient of HbA1c (steps 3 and 4 in Figure 1). The relation between HbA1c  
50 and systolic blood pressure was attenuated when measurement error was added to HbA1c,  
51 but not when measurement error was added to BMI. The association became stronger as  
52 measurement error was added solely to the confounding variable BMI. The reason for this  
53 effect is that, with increasing levels of measurement error on BMI, adjustment for the  
54 confounding due to BMI becomes less efficient and the effect estimate gets closer to the  
55 unadjusted estimate (1.13mmHg). Due to measurement error, a type of residual confounding  
56 is introduced. In the case of measurement error on HbA1c as well as BMI, both phenomena  
57  
58  
59  
60

1  
2  
3 play a role and may cancel each other out. In this study, measurement error on HbA1c  
4 seemed more influential than measurement error on BMI.  
5

6  
7 This example illustrates how a simple simulation study could provide insight into an  
8 important potential source of bias, namely measurement error. Here, we only considered  
9 classical measurement error, but simulations could easily be extended to incorporate more  
10 complex forms of measurement error. For example, the errors may not be drawn from a  
11 normal distribution with mean zero or may not be independent of all other variables  
12 considered. Instead, the mean of the distribution of errors may depend on the value of  
13 another variable in the model, e.g., error on BMI may depend on gender. Furthermore, non-  
14 normal distributions may be considered, or scenarios in which the variance of the errors  
15 depends on the true value of the measurement (heteroskedastic errors), among other  
16 possible extensions.  
17  
18  
19  
20

21 Finally, we note that researchers conducting small-scale simulation studies like the one  
22 presented here should reflect on the plausibility of the scenarios considered. For example,  
23 knowing whether it is realistic to assume that 50% of the total variance of HbA1c and BMI is  
24 due to measurement error (top right scenario in Figure 2) requires subject-matter  
25 knowledge.  
26  
27  
28

## 29 **6 CONCLUDING REMARKS**

30  
31 Just as randomised clinical trials form part of the evidence base for the choice of therapy in  
32 medical practice, simulation studies form part of the evidence base for statistical practice.  
33 Large-scale simulation studies allow assessment of the properties of complex estimation and  
34 inferential methods, and comparison of complex model building strategies under a variety of  
35 alternative assumptions and sample sizes.[5] They provide valuable support for decision-  
36 making regarding the choice of statistical methods to be used in a given real-life application  
37 and they are the cornerstone of the work on guidance for the design and analysis of the  
38 STRATOS initiative. They complement – rather than replace – the judgement of a trained  
39 expert (a data analyst in the choice of statistical methods, analogous to a physician in the  
40 choice of therapies). Increased computational power nowadays makes it possible to examine  
41 many possible simulation scenarios with different combinations of distributional parameters  
42 and assumptions. This partly addresses the main limitation of simulations, namely that they  
43 can never fully reflect the complexity of real data.  
44  
45  
46  
47  
48  
49

50 Let us again consider our analogy between simulation studies and clinical studies. The  
51 design and implementation of clinical studies should be left to teams of trained clinical  
52 researchers, but it is crucial for practitioners who want to practise evidence-based medicine  
53 to be able to read and understand the results of these clinical studies. Similarly, the design,  
54 implementation and reporting of complex simulations is still a subject of debate [3] and  
55 should be left to methodological statistical experts, but it is important for data analysts to be  
56 able to read and understand simulation studies in the literature (or perhaps to implement  
57 simple ones themselves). Armed with these skills, they will be better able to identify  
58  
59  
60

1  
2  
3 appropriate data analysis methods for their data and research questions, which will  
4 ultimately contribute to improved replicability of research results.  
5  
6  
7  
8

## 9 **ACKNOWLEDGMENTS**

10  
11 The authors thank Alethea Charlton for language corrections. The international  
12 STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative aims to  
13 provide accessible and accurate guidance for relevant topics in the design and analysis of  
14 observational studies (<http://stratos-initiative.org>). Members of simulation panel at the time of  
15 first submission: Michal Abrahamowicz, Harald Binder, Anne-Laure Boulesteix, Rolf  
16 Groenwold, Victor Kipnis, Tim Morris, Jessica Myers Franklin, Willi Sauerbrei, Pamela Shaw,  
17 Ewout Steyerberg, Ingeborg Waernbaum.  
18  
19  
20  
21  
22

## 23 **COMPETING INTERESTS**

24  
25 The authors declare no competing interests.  
26  
27  
28  
29  
30  
31

## 32 **PATIENT AND PUBLIC INVOLVEMENT STATEMENT**

33  
34 Patients or the public WERE NOT involved in the design, or conduct, or reporting, or  
35 dissemination plans of our research.  
36  
37  
38  
39

## 40 **AUTHOR CONTRIBUTIONS**

- 41  
42 - ALB initiated and coordinated the project, and wrote most of the manuscript.  
43  
44 - RG performed the example analysis and wrote the corresponding section.  
45  
46 - MA, HB, MB, RH, TM and JR critically revised the manuscript for important intellectual  
47 content.  
48  
49 - WS initiated and coordinated the project.  
50  
51  
52

53 All authors made substantial contributions to the manuscript's content, text and approved the  
54 final version.  
55  
56  
57  
58  
59  
60

## FUNDING STATEMENT

This project was partly funded by the German Research Foundation (DFG) with grants BO3139/4-3 to ALB and SA580/10-1 to WS. MA is a James McGill Professor at McGill University. His research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant 228203 and the Canadian Institutes of Health Research (CIHR) grant PJT-148946. TPM was funded by the UK MRC, grants MC\_UU\_12023/21 and MC\_UU\_12023/29.

## REFERENCES

- 1 Burton A, Altman DG, Royston P, et al. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006;25(24):4279-92.
- 2 Sigal MJ, Chalmers RP. Play It Again: Teaching statistics with Monte Carlo simulation. *Journal of Statistics Education* 2016;24(3):136-56.
- 3 Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019;38(11):2074-102.
- 4 Macleod MR, Michie S, Roberts I, et al. Biomedical research: increasing value, reducing waste. *Lancet* 2014;383:101-4.
- 5 Sauerbrei W, Abrahamowicz M, Altman DG, et al. STRENGTHENING analytical thinking for observational studies: the STRATOS initiative. *Statistics in Medicine* 2014;33(30):5413-32.
- 6 Rochon J, Gondan M, Kieser M. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology* 2012;12(1):81.
- 7 Lotterhos KE, Moore JH, Stapleton AE. Analysis validation has been neglected in the Age of Reproducibility. *PLoS Biology* 2018;16(12):e3000070-e70.
- 8 Brakenhoff TB, Van Smeden M, Visseren FL, et al. Random measurement error: why worry? An example of cardiovascular risk factors. *PLoS One* 2018;13(2):e0192298.
- 9 Abrahamowicz M, Beauchamp M-E, Fournier P, et al. Evidence of subgroup-specific treatment effect in the absence of an overall effect: is there really a contradiction? *Pharmacoepidemiology and Drug Safety* 2013;22(11):1178-88.
- 10 Boulesteix A-L, Wilson R, Hapfelmeier A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. *BMC Medical Research Methodology* 2017;17(1):138-38.
- 11 Boulesteix A-L, Binder H, Abrahamowicz M, et al. On the necessity and design of studies comparing statistical methods. *Biometrical Journal* 2018;60(1):216-18.
- 12 Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998;17(8):857-72.
- 13 Shmueli G. To Explain or to Predict? *Statistical Science* 2010;25(3):289-310.
- 14 De Bin R, Janitza S, Sauerbrei W, et al. Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics* 2016;72(1):272-80.
- 15 Boulesteix AL, Wright MN, Hoffmann S, et al. Statistical learning approaches in the genetic epidemiology of complex diseases. *Human Genetics* 2020;139:73-84.
- 16 De Bin R, Boulesteix AL, Benner A, et al. Combining clinical and molecular data in regression prediction models: insights from a simulation study. *Briefings in Bioinformatics* 2020.
- 17 Graf E, Schmoor C, Sauerbrei W, et al. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 1999;18(17-18):2529-45.
- 18 Boulesteix AL, Strobl C, Augustin T, et al. Evaluating microarray-based classifiers: an overview. *Cancer Informatics* 2008;6:77-97.

- 1  
2  
3 19 Hanczar B, Hua J, Dougherty ER. Decorrelation of the true and estimated classifier errors  
4 in high-dimensional settings. *EURASIP Journal on Bioinformatics and Systems*  
5 *Biology* 2007;38473.  
6  
7 20 Kent P, Jensen RK, Kongsted A. A comparison of three clustering methods for finding  
8 subgroups in MRI, SMS or clinical data: SPSS TwoStep Cluster analysis, Latent Gold  
9 and SNOB. *BMC Medical Research Methodology* 2014;14:113-13.  
10  
11 21 Coretto P, Hennig C. A simulation study to compare robust clustering methods based on  
12 mixtures. *Advances in Data Analysis and Classification* 2010;4(2):111-35.  
13  
14 22 Zipf G, Chiappa M, Porter KS, et al. National health and nutrition examination survey:  
15 plan and operations, 1999-2010. *Vital Health Stat 1* 2013(56):1-37.  
16  
17 23 Brakenhoff TB, Mitroiu M, Keogh RH, et al. Measurement error is often neglected in  
18 medical literature: a systematic review. *Journal of Clinical Epidemiology* 2018;98:89-  
19 97.  
20  
21 24 Carroll RJ, Ruppert D, Stefanski LA, et al. *Measurement error in nonlinear models: A*  
22 *modern perspective, Second Edition*: CRC Press, 2006.  
23

## FIGURE CAPTIONS

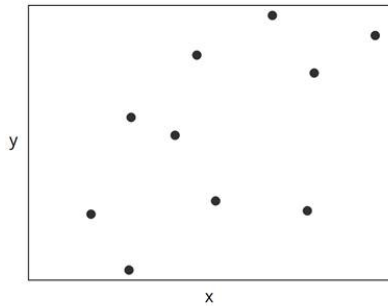
24  
25  
26 **Figure 1. Schematic illustration of the key steps of the simulation study described in**  
27 **section 5.**

28  
29  
30 **Figure 2. Estimates of the association between HbA1c levels and systolic blood**  
31 **pressure after adjustment for confounding by BMI under various simulation scenarios**  
32 **characterised by different levels of measurement error. Numbers represent effect**  
33 **estimates averaged over 1,000 simulation repetitions. Red shading represents low**  
34 **(averaged) estimates, blue shading represents high (averaged) estimates. Confidence**  
35 **intervals are omitted for clarity. See text for details.**  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

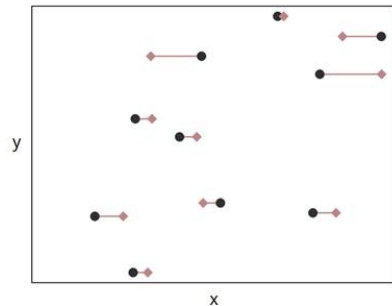


1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

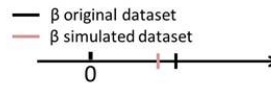
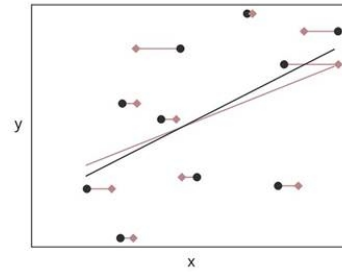
**1. Take original dataset**



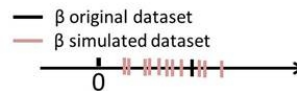
**2. Add random measurement error in x (by simulation)**



**3. Analyse the original dataset (1) and the simulated one with measurement error (2)**

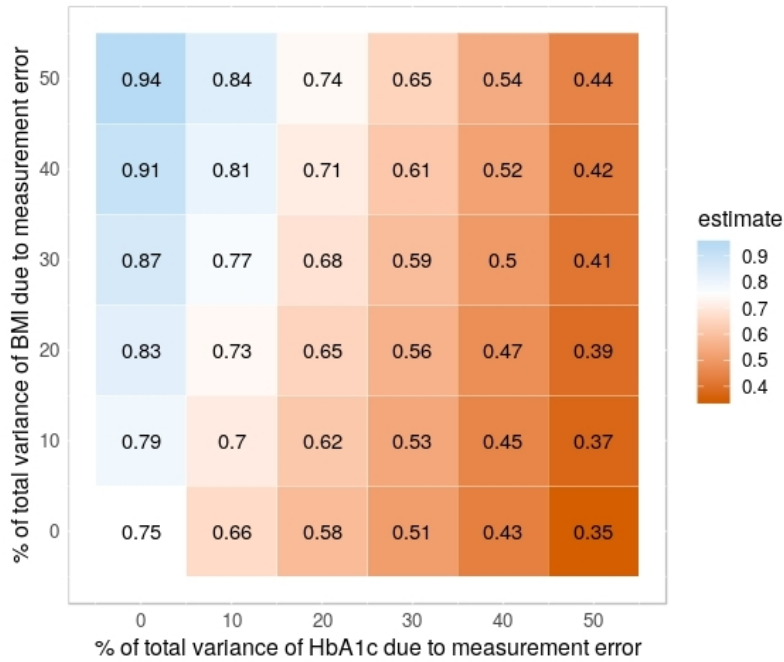


**Repeat for many simulated datasets**



Schematic illustration of the key steps of the simulation study described in section 5.

254x190mm (96 x 96 DPI)



Estimates of the association between HbA1c levels and systolic blood pressure after adjustment for confounding by BMI under various simulation scenarios characterised by different levels of measurement error. Numbers represent effect estimates averaged over 1,000 simulation repetitions. Red shading represents low (averaged) estimates, blue shading represents high (averaged) estimates. Confidence intervals are omitted for clarity. See text for details.

186x126mm (96 x 96 DPI)

```

1
2 # =====
3 # R CODE
4 # small scale simulation study to investigate impact of measurement error
5 # measurement error on (continuous) exposure and/or (continuous) confounding variable
6 # =====
7 #
8 # libraries:
9 library(Hmisc)
10 library(mice)
11 library(tidyverse)
12 #setwd("")
13 # =====
14 # set working directory:
15 # setwd("")
16 # =====
17 # The data can be downloaded in xpt form from https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015
18 nYear=2015
19 # read data:
20 d1 <- sasxport.get("DEMO_I.xpt")
21 d2 <- sasxport.get("BPX_I.xpt")
22 d3 <- sasxport.get("BMX_I.xpt")
23 d4 <- sasxport.get("GHB_I.xpt")
24 d5 <- sasxport.get("TCHOL_I.xpt")
25
26 d1.t <- subset(d1,select=c("seqn", "riagendr", "ridageyr"))
27 d2.t <- subset(d2,select=c("seqn", "bpxsy1"))
28 d3.t <- subset(d3,select=c("seqn", "bmxbmi"))
29 d4.t <- subset(d4,select=c("seqn", "lbggh"))
30 d5.t <- subset(d5,select=c("seqn", "lbdtsi"))
31
32 d <- merge(d1.t,d2.t)
33 d <- merge(d,d3.t)
34 d <- merge(d,d4.t)
35 d <- merge(d,d5.t)
36 # =====
37 # rename variables:
38 # RIAGENDR - Gender
39 # RIDAGEYR - Age in years at screening
40 # BPXSY1 - Systolic: Blood pres (1st rdg) mm Hg
41 # BMXBMI - Body Mass Index (kg/m**2)
42 # LBDTCSI - Total Cholesterol (mmol/L)
43 # LBXGH - Glycohemoglobin (%)
44
45 d$age <- d$ridageyr
46 d$sex <- d$riagendr
47 d$bp <- d$bpxsy1
48 d$bmi <- d$bmxbmi
49 d$HbA1C <- d$lbggh
50 d$chol <- d$lbdtsi
51 d$age[d$age<18] <- NA
52 # =====
53 # select complete cases:
54 dc <- cc(subset(d,select=c("age", "sex", "bmi", "HbA1C", "bp")))
55
56 # analysis:
57 summary(lm(bp ~ HbA1C + age + as.factor(sex), data=dc))
58 confint(lm(bp ~ HbA1C + age + as.factor(sex), data=dc))
59
60 summary(lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc))
61 confint(lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc))

```

```

1
2
3
4 # =====
5 # simulation of measurement error:
6 ref <- lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc)$coef[2]
7 n.sim <- 1e3
8 perc.me.exp <- seq(0,.5,.1)
9 perc.me.conf <- seq(0,.5,.1)
10 scenarios <- expand.grid(perc.me.exp,perc.me.conf)
11 var.exp <- var(dc$HbA1C)
12 var.conf <- var(dc$bmi)
13 n <- dim(dc)[1]
14 beta.hat <- matrix(ncol=dim(scenarios)[1], nrow=n.sim)
15
16 for (k in 1:n.sim){
17   print(k)
18   set.seed(k)
19   for (i in 1:dim(scenarios)[1]){
20     var.me.exp <- var.exp*scenarios[i,1]/(1-scenarios[i,1])
21     var.me.conf <- var.conf*scenarios[i,2]/(1-scenarios[i,2])
22     dc$HbA1C.me <- dc$HbA1C + rnorm(dim(dc)[1], 0, sqrt(var.me.exp) )
23     dc$bmi.me <- dc$bmi + rnorm(dim(dc)[1], 0, sqrt(var.me.conf) )
24     beta.hat[k,i] <- lm(bp ~ HbA1C.me + age + bmi.me + as.factor(sex), data=dc)$coef[2]
25   }
26 }
27 # =====
28 # create figure:
29 tot.mat <- cbind(100*scenarios,apply(beta.hat,2,mean))
30 colnames(tot.mat) <- c("me.exp","me.conf","estimate")
31
32 FIGURE <- ggplot(tot.mat, aes(me.exp, me.conf)) +
33   geom_tile(color="white",aes(fill = estimate)) +
34   geom_text(aes(label = round(estimate, 2))) +
35   scale_fill_gradient2(low="#D55E00",mid="white",high = "#56B4E9", midpoint=ref) +
36   labs(x=paste("% of total variance of HbA1c due to measurement error"),
37        y=paste("% of total variance of BMI due to measurement error")) +
38   coord_equal()+
39   scale_y_continuous(breaks=unique(tot.mat[,1]))+
40   scale_x_continuous(breaks=unique(tot.mat[,1]))+
41   theme(panel.background = element_rect(fill='white', colour='grey'),
42         plot.title=element_text(hjust=0),
43         axis.ticks=element_blank(),
44         axis.title=element_text(size=12),
45         axis.text=element_text(size=10),
46         legend.title=element_text(size=12),
47         legend.text=element_text(size=10))
48
49 FIGURE
50 # savePlot("Figure_STRATOS.tif", type="tif")
51 # =====
52 # END OF R CODE
53 # =====
54
55
56
57
58
59
60

```