PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	An introduction to statistical simulations in health research
AUTHORS	Boulesteix, Anne-Laure; Groenwold, Rolf; Abrahamowicz, Michal; Binder, Harald; Briel, Matthias; Hornung, Roman; Morris, Tim; Rahnenführer, Jörg; Sauerbrei, Willi

VERSION 1 – REVIEW

REVIEWER	Lisa Lix University of Manitoba, Canada
REVIEW RETURNED	29-Jul-2020

GENERAL COMMENTS	This manuscript aims to provide an introduction to the use of simulation studies in health research; the purpose is to "demonstrate that simulation studies are an invaluable tool" for making decisions about the choice of statistical methods for health research. The authors have substantial expertise in the design and analysis of observational studies through the STRATOS (STRengthening Analytical Thinking for Observational Studies) initiative. This article is intended to be published in the "Communications" section of BMJ Open. As such, the criteria provided to reviewersfor evaluation of submissions include: (a) appeal to a broad audience that includes patients, researchers, policy makers, health professionals, and doctors of all disciplines, (b) offers novel insights that have not been considered in existing literature, (c) aids in decision making, and (d) demonstrate one or more of the following values: transparency, openness, collaboration, innovation, reproducibility, patient/ public involvement, improving peer review and journal best practice, and reducing research waste. Main Comments This manuscript only partially addresses the review criteria: • Audience: The manuscript is targeted to researchers and data analysts; it is unlikely to have broad relevance for other stakeholder groups identified in the review criteria unless it is revised substantially. There is brief mention of practitioners in the conclusions section with reference to having skills in interpreting and understanding simulation studies. However, this target audience is not identified in the introduction. Moreover, if practitioners are a target audience, then they authors must consider whether to include information in this manuscript about best practices in

reporting on simulation studies; this type of information could be useful to practitioners in identifying high quality simulation studies. The cited paper by Burton et al., 2006, Statist. Med. 2006; 25:4279–4292, provides a useful guide to best practices, and this topic could be expanded upon in this manuscript.

• Decision Making: Simulation techniques can aid in decision making about the selection of statistical methods for complex, observational studies, and the authors have addressed this issue in the manuscript.

• Novel Insights: In the summary of strengths and limitations, the authors allude to the potential value of statistical simulations in the conduct of pandemic-related studies, but do not address this topic in the main body of the manuscript. The topical nature/novel aspects of the manuscript for the current health research environment are not described in the manuscript.

• Demonstrate Values: Well-designed simulation studies about statistical methods can improve the reproducibility of research. This can be accomplished by identifying the data-analytic conditions under which a statistical method will and will not perform optimally (e.g., the conditions under which a statistical test will produce empirical Type I error rates that are robust to departures from derivational assumptions that underlie these tests). The contributions of well-designed statistical simulations to contribute to reproducibility requires further discussion within the main body of the manuscript. Other values such as transparency and reducing research waste, which could be addressed through statistical simulations, could be expanded upon within the discussion section.

Additional Comments

I have a number of additional comments:

(a) From the outset, I would recommend that the authors emphasize the value of methods-focused simulation studies for improving the scientific rigour of analyses of complex observational studies. The two opening paragraphs of this manuscript focus more generally on "health research". The strengths of these STRATOS authors clearly lies within the domain of observational studies and appropriate statistical methods to analyze data arising from observational studies. It is not until the third paragraph that the authors discuss observational studies; this emphasis should occur earlier on.

(b) The manuscript would benefit greatly from a diagram that illustrates the sequence of steps in a simulation study. Visualizations are often a useful tool to convey complex concepts to diverse audiences.

(c) Section 3: Table 1 describes a scenario for the combination of A-B-C-D, but this scenario is not

discussed in the manuscript. Also, the last column of Table 1 may benefit from a different title
and explanation within the body of the manuscript, as it is not clear what "aim" refers to.
Furthermore, Type I error rates of a well-performing statistical test should be close to the
nominal level of significance, not "low". If the Type I error rate of a statistical test is very close to
0, then it is conservative, which could, in fact, affect statistical power. Bradley's 1978 paper
(British Journal of Mathematical and Statistical Psychology, 31, 144-152) on robustness provides
guidance for interpreting the performance of statistical tests.
(d) The numeric example would benefit substantially from organizing the information provided
around the key steps in a simulation study described in section 4. Here again, a diagram may be
helpful to guide the reader through the steps in this simulation example.

REVIEWER	Julius Sim Keele University, UK
REVIEW RETURNED	31-Aug-2020

GENERAL COMMENTS	An interesting and clearly written paper that will be helpful to readers seeking a clear and non-technical introduction to the field of simulation studies. I have a few suggestions, mostly very minor: Page 3, line 46: the link to the coronavirus crisis seems a little tenuous! I agree that the public is currently exposed to more statistical information, and associated methods, but I assume that your intended readership is clinicians and non-specialist researchers, rather than the general public. Page 4 line 11: 'applications' is a little unclear. I presume this means 'in applied settings', rather than applications in the sense of, e.g., computing programs. Page 4 line 14: I would describe relationships, rather than effects, as (non)linear in regression. Indeed, I think the word 'effects' is debatable here, as it suggests a causal process, and cause-effect relationships cannot necessarily be inferred from regression analyses. Page 4 line 36: insert 'their' before 'simulations'. Page 5 line 36: change 'on' to 'of'. Page 6 line 8: sample size is clearly an issue for the appropriateness of certain analyses, and the reliability of their outputs, but I wonder if it would normally be described as an 'assumption' of such analyses? Perhaps it is more often a factor that may influence the satisfaction of certain assumptions, rather than being an assumption per se. Page 6 line 60: 'break' is a rather unusual choice of word here. Page 7 line 7: it is rather unusual to include a hyphen in '95% confidence interval'. Page 8 line 44: I would question the appropriateness of using the terms 'independent variable' and 'dependent variable' in relation to regression, as they imply possibly unwarranted assumptions (e.g. that the Y variable does indeed 'depend' on the X variables, and that the X variables are independent). I would suggest alternative terms, like 'predictor' and 'outcome variable' (though I realize that you may have deliberately reserved these for when later

discussing regression as a predictive approach, so you could possibly use other alternatives).

Page 8 line 15: as this paper is directed at non-specialist methodologists, the use of terms like sensitivity and specificity might be confusing. It might help to say that they are being used by analogy with their use in diagnostic statistics.

Page 10 line 25: do you mean 'e.g.' rather than 'i.e.'?

Page 10 line 53; I would suggest starting a new sentence with 'However,' – otherwise the sentence has two semi-colons and is rather complex.

Page 10 line 57: I don't think 'stability' is adequately defined for the non-specialist reader.

Table 1: in the column 'evaluation criteria', I would be explicit and say 'coverage of confidence interval', particularly in relation to B, as CIs are not mentioned in this row. The row A-B-C-D is unclear. I presume that these criteria refer to any combination of A, B, C and D? I don't think model convergence has been defined in the paper, and the non-specialist reader may not understand the term.

Page 12 line 8: maybe 'relevant to' rather than 'relevant for'.

Page 12 line 36: in view of previous comments about the limitations of using real datasets, the phrase 'using real datasets as a basis' may need fuller explanation if it is not to confuse readers.

Page 12 line 42: in what sense would their complexity be 'arbitrary'?

Page 13 line 42: I don't think parameters per se can be described as 'true'. Maybe say 'true values of its parameters'.

Page 14 line 58: I would suggest using small 'n' when denoting sample size, as in sampling theory large 'N' is often reserved for the population size (if known).

Page 14 line 24: why 'On the other hand'? The implied contrast with the preceding sentence is not clear.

Page 14 lines 41 and 42: I would remove the commas after 'test' and 'distribution', as they convey a sense that I don't think you intend.

Page 14 line 46: be a little more explicit as to what these scenarios correspond to.

Page 15 line 6: I think this would read more clearly if you omitted the comma after 'situations'.

Page 15 line 36: the idea of an ethically unacceptable setting is rather unclear. Does this mean a case in which the collection of the necessary first-hand data would require creating a situation that would raise ethical problems?

Page 15 line 38: 'randomisation' usually denotes random allocation. Is that the sense in which you are using it here, or do you mean random sampling?

Page 15 line 36: suggest 'performance' rather than 'performances'.

Page 18 line 11: suggest 'increased' rather than 'increasing', as I think you are using the word adjectivally rather as denoting an action.

Page 18 line 21: 'practise' rather than 'practice'.

References: there is inconsistency in the (non)abbreviation of journal titles and in capitalization in article titles (and the journal title in reference 22).

Figure: I think the meaning of the shading in this figure could be made clearer in the legend.

Reviewer 1 (Lisa Lix):

This manuscript aims to provide an introduction to the use of simulation studies in health research; the purpose is to "demonstrate that simulation studies are an invaluable tool" for making decisions about the choice of statistical methods for health research. The authors have substantial expertise in the design and analysis of observational studies through the STRATOS (STRengthening Analytical Thinking for Observational Studies) initiative.

Thanks!

This article is intended to be published in the "Communications" section of BMJ Open. As such, the criteria provided to reviewers for evaluation of submissions include: (a) appeal to a broad audience that includes patients, researchers, policy makers, health professionals, and doctors of all disciplines, (b) offers novel insights that have not been considered in existing literature, (c) aids in decision making, and (d) demonstrate one or more of the following values: transparency, openness, collaboration, innovation, reproducibility, patient/ public involvement, improving peer review and journal best practice, and reducing research waste.

Main Comments

This manuscript only partially addresses the review criteria:
• **Audience:** The manuscript is targeted to researchers and data analysts; it is unlikely to have broad relevance for other stakeholder groups identified in the review criteria unless it is revised substantially. There is brief mention of practitioners in the conclusions section with reference to having skills in interpreting and understanding simulation studies. However, this target audience is not identified in the introduction. Moreover, if practitioners are a target audience, then they authors must consider whether to include information in this manuscript about best practices in reporting on simulation studies; this type of information could be useful to practitioners in identifying high quality simulation studies. The cited paper by Burton et al., 2006, Statist. Med. 2006; 25:4279–4292, provides a useful guide to best practices, and this topic could be expanded upon in this manuscript.

We must admit that the targeted audience primarily consists of clinical researchers rather than practitioners. However, we think that, from a general point of view, practitioners should become more aware of the impact of the choice of statistical methods on the conclusions of the studies they read. Following the reviewer's recommendation, we added a sentence in the introduction to draw attention to this aspect:
"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."

• **Decision Making:** Simulation techniques can aid in decision making about the selection of statistical methods for complex, observational studies, and the authors have addressed this issue in the manuscript.
• **Novel Insights:** In the summary of strengths and limitations, the authors allude to the potential value of statistical simulations in the conduct of pandemic-related studies, but do not address this topic in the main body of the manuscript. The topical nature/novel aspects of the manuscript for the current health research environment are not described in the manuscript.

The principles of statistical simulation studies, and the general methodological challenges involved in the assessment and interpretation of their results have, to our knowledge, not previously been

introduced to clinical researchers . In this sense, our approach provides novel insights and uses a novel approach to explain the relevant concepts in a "nontechnical" way. We believe that clinical researchers should at least know the basics of statistical simulation studies and we make an original contribution to the literature for this purpose. This is now stated more explicitly in the introduction:

"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."

• **Demonstrate Values:** Well-designed simulation studies about statistical methods can improve the reproducibility of research. This can be accomplished by identifying the data-analytic conditions under which a statistical method will and will not perform optimally (e.g., the conditions under which a statistical test will produce empirical Type I error rates that are robust to departures from derivational assumptions that underlie these tests). The contributions of well designed statistical simulations to contribute to reproducibility requires further discussion within the main body of the manuscript. Other values such as transparency and reducing research waste, which could be addressed through statistical simulations, could be expanded upon within the discussion section.

Thanks for mentioning this very important point. We now emphasize these aspects both in the introduction and in the discussion.

"The use of inappropriate statistical methods, that affects a vast spectrum of research activities, contributes to the replication crisis that has drawn increasing attention in recent years; see for example the Lancet series "Increasing value, reducing waste". 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."

"Armed with these skills, they will be better able to identify appropriate data analysis methods for their data and research questions, which will ultimately contribute to improved replicability of research results."

Additional Comments

I have a number of additional comments:

(a) From the outset, I would recommend that the authors emphasize the value of methods-focused simulation studies for improving the scientific rigour of analyses of complex observational studies. The two opening paragraphs of this manuscript focus more generally on "health research". The strengths of these STRATOS authors clearly lies within the domain of observational studies and appropriate statistical methods to analyze data arising from observational studies. It is not until the third paragraph that the authors discuss observational studies; this emphasis should occur earlier on.

Even though the focus of the STRATOS initiative is on observational studies, as the reviewer correctly recognized, our introduction to simulation studies may also be useful for researchers from other fields such as clinical trials, where the choice of statistical methods is also crucial (note that the expertise of the team of authors is not limited to observational studies). We thus decided to keep the original formulation, but now explicitly state that simulation studies are useful to a broad field of studies, to avoid misunderstandings:

"for example, data from observational studies or from clinical trials"

(b) The manuscript would benefit greatly from a diagram that illustrates the sequence of steps in a simulation study. Visualizations are often a useful tool to convey complex concepts to diverse audiences.

Thanks for this suggestion. We now include an additional table (table 2) which gives an overview of the key features of simulation studies – in general and for our example simulation, as well as a new figure (figure 1) displaying the different steps of our example simulation schematically.

(c) Section 3: Table 1 describes a scenario for the combination of A B-C-D, but this scenario is not discussed in the manuscript.

This row actually includes the evaluation criteria that are relevant to all settings A, B, C and D rather than to a specific one. To clarify, we replaced "A-B-C-D" by "all settings" in the table.

Also, the last column of Table 1 may benefit from a different title and explanation within the body of the manuscript, as it is not clear what "aim" refers to. Furthermore, Type I error rates of a wellperforming statistical test should be close to the nominal level of significance, not "low". If the Type I error rate of a statistical test is very close to 0, then it is conservative, which could, in fact, affect statistical power. Bradley's 1978 paper (*British Journal of Mathematical and Statistical Psychology,* 31, 144- 152) on robustness provides guidance for interpreting the performance of statistical tests.

We changed the column title to "target value" and clarified its meaning by adding the following sentence:

"The last column indicates which values the considered evaluation criterion takes if the investigated method is good."

Regarding the type I error, we completed the entry of the last column as "close to and not greater than nominal value α".

(d) The numeric example would benefit substantially from organizing the information provided around the key steps in a simulation study described in section 4. Here again, a diagram may be helpful to guide the reader through the steps in this simulation example.

Thanks for this suggestion. We now include an additional table (table 2) which gives an overview of the key features of simulation studies – in general and for our example simulation, as well as a new figure (figure 1) that schematically displays the different steps of our example simulation.

Reviewer 2 (Julius Sim):

An interesting and clearly written paper that will be helpful to readers seeking a clear and non-technical introduction to the field of simulation studies.

Thanks!

I have a few suggestions, mostly very minor:

Page 3, line 46: the link to the coronavirus crisis seems a little tenuous! I agree that the public is currently exposed to more statistical information, and associated methods, but I assume that your intended readership is clinicians and non-specialist researchers, rather than the general public.

The article summary, in which we referred to the current pandemic, has been removed.

Page 4 line 11: 'applications' is a little unclear. I presume this means 'in applied settings', rather than applications in the sense of, e.g., computing programs.´

Done!

Page 4 line 14: I would describe relationships, rather than effects, as (non)linear in regression. Indeed, I think the word 'effects' is debatable here, as it suggests a causal process, and cause-effect relationships cannot necessarily be inferred from regression analyses.

Done!

Page 4 line 36: insert 'their' before 'simulations'. Done!

Page 5 line 36: change 'on' to 'of'.

Done!

Page 6 line 8: sample size is clearly an issue for the appropriateness of certain analyses, and the reliability of their outputs, but I wonder if it would normally be described as an 'assumption' of such analyses? Perhaps it is more often a factor that may influence the satisfaction of certain assumptions, rather than being an assumption per se.

We removed "large sample size" to avoid misunderstandings.

Page 6 line 60: 'break' is a rather unusual choice of word here.

We replaced "break" by "fail".

Page 7 line 7: it is rather unusual to include a hyphen in '95% confidence interval'.

Done!

Page 8 line 44: I would question the appropriateness of using the terms 'independent variable' and 'dependent variable' in relation to regression, as they imply possibly unwarranted assumptions (e.g. that the Y variable does indeed 'depend' on the X variables, and that the X variables are independent). I would suggest alternative terms, like 'predictor' and 'outcome variable' (though I realize that you may have deliberately reserved these for when later discussing regression as a predictive approach, so you could possibly use other alternatives).

We changed "dependent variable" to "outcome variable" and "independent variable" to "covariate" throughout the manuscript, while also mentioning the alternative phrases once.

Page 8 line 15: as this paper is directed at non-specialist methodologists, the use of terms like sensitivity and specificity might be confusing. It might help to say that they are being used by analogy with their use in diagnostic statistics.

We added the following explanations:

"Regarding (i), it is good to have high sensitivity (i.e., selecting most/all variables with effects, this is analogous to detecting most/all diseased patients in a diagnostic study) as well as high specificity (i.e., not selecting variables without an effect, analogous to correctly identifying participants without disease)."

Page 10 line 25: do you mean 'e.g.' rather than 'i.e.'?

No, we meant "e.g.," in the sense of "for example".

Page 10 line 53; I would suggest starting a new sentence with 'However,' – otherwise the sentence has two semi-colons and is rather complex.

Done!

Page 10 line 57: I don't think 'stability' is adequately defined for the non-specialist reader.

We added the explanation "i.e., robustness against small changes in the data":

"Real data may be used to assess aspects such as stability (i.e., robustness against small changes in the data)"

Table 1: in the column 'evaluation criteria', I would be explicit and say 'coverage of confidence interval', particularly in relation to B, as CIs are not mentioned in this row.

Done!

The row A-B-C-D is unclear. I presume that these criteria refer to any combination of A, B, C and D? I don't think model convergence has been defined in the paper, and the non-specialist reader may not understand the term.

This row actually includes the evaluation criteria that are relevant to all settings A, B, C and D rather than to a specific one. To clarify, we replaced "A-B-C-D" by "all settings" in the table. Regarding convergence, we reformulated this item as "success of the computation (e.g., convergence)", since a computation may fail for other reasons than lack of convergence (e.g., memory issues).
Page 12 line 8: maybe 'relevant to' rather than 'relevant for'. Done!

Page 12 line 36: in view of previous comments about the limitations of using real datasets, the phrase 'using real datasets as a basis' may need fuller explanation if it is not to confuse readers.

We replaced this phrase by "by using real datasets following as a starting point (see section 5 for an example)".

Page 12 line 42: in what sense would their complexity be 'arbitrary'?

The term "arbitrarily" was indeed misleading. We removed it.

Page 13 line 42: I don't think parameters per se can be described as 'true'. Maybe say 'true values of its parameters'.

We replaced "its true parameters" by "the true values of its parameters" as suggested by the reviewer.

Page 14 line 58: I would suggest using small 'n' when denoting sample size, as in sampling theory large 'N' is often reserved for the population size (if known).

 Done!

Page 14 line 24: why 'On the other hand'? The implied contrast with the preceding sentence is not clear.

We agree and removed "On the other hand".

Page 14 lines 41 and 42: I would remove the commas after 'test' and 'distribution', as they convey a sense that I don't think you intend.

 Done!

Page 14 line 46: be a little more explicit as to what these scenarios correspond to.

We added the following sentence:

"Scenarios may differ, among other ways, in the sample size, the true distributions of the considered variables (normal, uniform, exponential, etc.), the values of parameters such as means or variances, the correlation structure of the variables or the presence of outliers."

Page 15 line 6: I think this would read more clearly if you omitted the comma after 'situations'.

Done!

Page 15 line 36: the idea of an ethically unacceptable setting is rather unclear. Does this mean a case in which the collection of the necessary first-hand data would require creating a situation that would raise ethical problems?

We think that this aspect is not so important and thus does not deserve a long explanation, so we removed it.

Page 15 line 38: 'randomisation' usually denotes random allocation.

Is that the sense in which you are using it here, or do you mean random sampling?

Yes, it was what we meant. To make this unequivocal, we replaced "randomisation" by "randomised trial data".

 Page 15 line 36: suggest 'performance' rather than 'performances'.

 Done!

Page 18 line 11: suggest 'increased' rather than 'increasing', as I think you are using the word adjectivally rather as denoting an action.

Done!

Page 18 line 21: 'practise' rather than 'practice'.

Done!

References: there is inconsistency in the (non)abbreviation of journal titles and in capitalization in article titles (and the journal title in reference 22).

Done!

Figure: I think the meaning of the shading in this figure could be made clearer in the legend.

We added the sentence:

"Red shading represents low (averaged) estimates, blue shading represents high (averaged) estimates."

Furthermore, we also amended the text:

"Figure 2 shows the impact of measurement error on HbA1c and/or BMI on the estimate of the regression coefficient of HbA1c."

VERSION 2 – REVIEW

REVIEWER	Lisa Lix University of Manitoba, Canada
REVIEW RETURNED	06-Nov-2020

GENERAL COMMENTS	I have no further comments on this manuscript.

REVIEWER	Julius Sim Keele University, UK
REVIEW RETURNED	19-Oct-2020

GENERAL COMMENTS	Many thanks for responding to my comments.