OBSTETRICS & GYNECOLOGY

NOTICE: This document contains correspondence generated during peer review and subsequent revisions but before transmittal to production for composition and copyediting:

- Comments from the reviewers and editors (email to author requesting revisions)
- Response from the author (cover letter submitted with revised manuscript)*

**The corresponding author has opted to make this information publicly available.*

Personal or nonessential information may be redacted at the editor's discretion.

Questions about these materials may be directed to the *Obstetrics & Gynecology* editorial office: obgyn@greenjournal.org.

Date: Dec 20, 2019
To: "Kartik Kailas Venkatesh" [redacted]
From: "The Green Journal" em@greenjournal.org
Subject: Your Submission ONG-19-2156

RE: Manuscript Number ONG-19-2156

Machine learning and statistical models to predict postpartum hemorrhage

Dear Dr. Venkatesh:

Your manuscript has been reviewed by the Editorial Board and by special expert referees. Although it is judged not acceptable for publication in Obstetrics & Gynecology in its present form, we would be willing to give further consideration to a revised version.

If you wish to consider revising your manuscript, you will first need to study carefully the enclosed reports submitted by the referees and editors. Each point raised requires a response, by either revising your manuscript or making a clear and convincing argument as to why no revision is needed. To facilitate our review, we prefer that the cover letter include the comments made by the reviewers and the editor followed by your response. The revised manuscript should indicate the position of all changes made. We suggest that you use the "track changes" feature in your word processing software to do so (rather than strikethrough or underline formatting).

Your paper will be maintained in active status for 21 days from the date of this letter. If we have not heard from you by Jan 10, 2020, we will assume you wish to withdraw the manuscript from further consideration.

REVIEWER COMMENTS:

Reviewer #1: General comments:
Overall this paper addresses a timely and important issue in obstetrics through application of novel methods. The main challenge the authors of this paper face is the task of communicating complex statistical and machine learning methods to the average practicing Ob/Gyn in an approachable and informative way. Towards this end, the authors spend a considerable portion of the paper explaining their methods and teaching the reader about how these models are developed, interpreted and evaluated. Overall, the authors accomplish this task, though I think a few points could be further clarified which I will delineate below.

Introduction:
For the background portion of the introduction, I would consider adding in a short statement about how accurate prediction of adverse events such as PPH can allow for improved clinical outcomes. Additionally, the authors mention rates of transfusion though do explicitly state the utility of their models for predicting transfusion rates.

The third paragraph of the introduction contrasting traditional statistical models and machine learning is a bit of an abrupt transition. A statement such as "Current methods for predicting PPH are based on risk stratification methods. However improved predictive ability could be achieved through application of traditional statistical models and machine learning. Recent advances in computer science have led to the development of machine learning…which has superior predictive ability over traditional statistical methods." It is important to the paper to situate these two approaches (statistical and machine learning models) in relation to current, less individualized risk stratification approaches.

Excellent explanation of machine learning!

Methods:
Consider changing the description of the CSL database to present tense (vs past tense).

For the first paragraph on page 9, consider clarifying that logistic regression is the first of the two statistical models. The remainder of this paragraph includes a very brief explanation of the two machine learning models and includes several terms that would be unfamiliar to the average reader. I would consider either further explaining these terms ("bootstrap samples", "training data," "ensembles") or perhaps leaving these explanations out entirely and making a more general statement that two types of machine learning models were created as it is not entirely necessary that the reader understand the models—only the conclusion that such models can be made and have good predictive ability.

In the last paragraph on page 9 (begins "Measures, such as sensitivity…") explaining how model performance is assessed the authors do an excellent job of educating the reader and positioning them to understand what follows.

The last paragraph on page 12 (begins "Figure 3 demonstrates…") and the first paragraph on page 13 (begins "In the decision curve analysis…") it is unclear to me what the following means: "This model assigned an accurate probability of post partum hemorrhage when prediction ranged from 0 to 70-80%." Perhaps it is clear to those more familiar with these types of tests/statistics.

For the last paragraph of the results section on page 13 (begins "We were unable to compare our models…") I would consider additional clarification as this paragraph addresses the secondary aim of the study. A suggestion "We were unable to compare our models to current risk stratification strategies as outlined by CMQCC (i.e., low, medium, and high risk for post partum hemorrhage) as only two thirds of the variables included in CMQCC guidelines were identified by our models and many of the variables included in CMQCC (such as uterine fibroids, known bleeding disorder or coagulopathy) were not routinely assessed on labor admission and therefore were not considered as candidate variables for our models."

Discussion

While the authors clearly state that machine learning models outperformed the statistical models, I would consider including a more robust discussion of what level of discrimination is clinically relevant. This is particularly salient as they immediately state that as statistical models are more easily incorporated into practice in EMRs that perhaps statistical models, rather than machine learning models, would be a reasonable alternative as they performed reasonably well in predictive ability.

The authors make a general statement that "any model will need to be prospectively assessed in local contemporary cohorts of pregnant women across regional settings." Do the authors intend to further prospectively validate these models?

Reviewer #2: The authors aim to construct models to predict PPH using both machine learning and more traditional statistical modeling. I have the following comments regarding the manuscript:

Abstract

1. The Conclusion about these models allowing providers to transfer women to appropriate levels of maternal care goes way beyond what can be concluded from this particular project.

Intro

1. Consider deleting the possible secondary aim of comparison to the CMQCC risk stratification. We all brainstorm research ideas and then go about determining if they are feasible. Since this was found not to be feasible, it does not need to be included in the manuscript other than perhaps a mention in the discussion of the limitations of the CSL dataset and inability to compare to other risk stratification systems.

Methods

1. The authors constructed two machine learning models, and two more traditional statistical models. Was one of the objectives to compare these different methods for predicting PPH?

2. Can the authors clarify how lasso regression provides a "different method" of identifying the final variables in the model compared to standard logistic regression?

3. The paragraph about the decision not to use sensitivity, specificity, etc could be significantly pared down. Just tell the reader what you did to analyze the data.

4. Consider first saying that decision curve analysis (DCA) was used to quantify the net benefit of the models, and then further describe DCA. Otherwise the authors are well into the description of DCA before the reader knows what it was used for in this manuscript.

5. How did the authors deal with missing data?

Results

1. A 2% rate of blood transfusion for births WITHOUT PPH seems quite high. Do the investigators suspect a data collection error? Please explain.

2. The authors compare the models based on their discriminatory capability and report higher or lower discriminatory capacity. Can the c-statistic be compared statistically?

3. Would recommend being very precise about when you are using individual hospitals versus sites. For example text about figure 3 says you are comparing sites and the legend says hospitals. Would double check this throughout.

4. Why is the upper limit of prediction listed at 70-80%? Is it because prediction was broken into deciles for analysis? If so, this should be reported in the methods.

5. The idea of net benefit is not presented comprehensively until the legend of Figure 4. Why aren't these equations in the methods?

6. Was the prenatal diagnosis of macrosomia in CSL by Leopold's or ultrasound? Was it universally assessed? I doubt it since the CSL data were pulled retrospectively.

7. Figure 3. Have the authors looked in detail at data from hospital 41 to ensure that it does not appear to be flawed/missing, etc? It is clearly out of line when compared to other hospitals.

8. Figure 5. Why abbreviate the predictors on the y-axis with non-intuitive abbreviations?

Discussion

1. Last paragraph is appropriately tempered to convey potential down the road uses. Need to tone down the idea of transfer to higher level of care based on these models.

2. 50 references is a lot for original research. Can the list be shortened to critical references?

3. If the actual models are not provided, how can the reader use them or validate them?

Reviewer #3: In this manuscript, the authors describe the application of a relatively new approach to analyze data regarding the common obstetric complication of postpartum hemorrhage (PPH). They use a large, publicly available dataset to build and compare models for prediction of individual patient PPH risk based on information available at the time of admission. Although this has been done previously with other datasets, the major novelty of this project is the use of two different machine learning approaches, in addition to statistical models, and compared their accuracy for prediction in their validation process. Indeed, one of the major strengths of this manuscript is that it could be one of the first publications in our field to demonstrate the potential for machine learning to create highly accurate predictive models for important individual patient outcomes. This would presumably provide superior performance to current guidelines which are based in some part on expert opinion without validation. Another strength is the considerations elaborated in the Discussion regarding limitations and further developments needed before an optimal model is ready for clinical use.

The methods section contains a very complex, extended description of what the models entailed and the rationale for the use of four different models. The descriptions of how the analyses should be interpreted is very much appreciated. Full understanding of the methods will be beyond most readers. Yet, it can be appreciated that data supports the conclusions reached.

The manuscript does specifically state that the prediction is based on factors present at the time of patient admission. There are recognized intrapartum risk factors that can affect the risk of PPH, such as prolonged labor/stimulation with oxytocin and intrapartum chorioamnionitis, which are therefore not included. With the predictive accuracy described, it is hard to imagine that the C-statistic could get much higher; but presumably a model which could incorporate important developments after admission would have more clinical value at the time of delivery or after. And as stated in the Introduction, to objective was to develop and validate a predictive model for PPH, not limited to those available at the time of admission. An argument can be made for performing the analysis upon admission (for the purpose of This could be discussed as part of limitations.

Comments:

1. Methods—it is not clear from the text that the elements listed in Appendix Table 3 are the elements used in the modeling. This table is difficult to read because of the way it is formatted. Consider indent every other line with the Missing n so that the actual elements appear lined up to the left margin.

2. Results—in the second paragraph of the results, 14 variables are listed in the first sentence. The second sentence contains another 16 or so variables are listed and described as associated with PPH. Presumably we are to infer that those listed in the first sentence are not associated with PPH. It is not clear why you list this many of them, but not all 55.

3. Results—in the fourth paragraph, it would be helpful if it were explained that this analysis refers to the data presented in the second column of table 1.

4. Figure 3—the legend is very helpful in understanding what is being done, but more explanation in the text would be helpful as to how the accuracy probably was assigned, and what the meaning of this is.

5. Figure 4—similarly, please describe in the text what is seen in the figure which allows for the conclusion being made.

STATISTICAL EDITOR'S COMMENTS:

1. Generalizability: As can be seen in Table S3 (some of this information should be shown more prominently in main text), those with missing data for EBL (and therefore omitted from the model analyses) were ~ 1/3 of the original cohort, with 67.0% available for vaginal and 65.8% for cesarean deliveries. This proportion of missing data is a major limitation to general application of these conclusions. Specifically for the model building, some important variables had high rates of missing data (eg, pre-pregnancy wgt (29.4% missing overall, with 45.8% missing among the PPH subset). Proportions less, but still important for another important variable, adm wgt: 15.7% missing overall with 23.0% missing among the PPH cohort. Also, for macrosomia, the overall cohort had 44.7% missing data, while for PPH, the missing data rate was 5.6%. These three variables were cited in Fig 5 as comprising the highest importance values for the nominally best prediction model for PPH. Even if these were imputed (pg 10, last line, ref (35)), the proportion of missing data seriously limits imputation and therefore how this model faithfully reflects the population studied.

2. General: Machine learning would be new topic for most readers, so there should be a short primer on its methods (could be supplemental material) with concise summary in main text.

3.It is unclear from Methods whether the same variables were retained in each of the 4 models evaluated. If so, should so state. If not, then should construct a Table listing the retained variables in each model in rank order, with the relative importance of each to the final prediction. Also, it seems from the general results that cesarean delivery is associated with higher risk of PPH than vaginal delivery. Why was CD not included as a potential variable, or analyzed as a separate cohort, since > 90% of the primary outcome was associated with CD? That variable alone would have high discriminatory value for prediction of PPH.

4. Table 1: This Table needs much more elaboration. The model results (C-statistic with CIs) from the first phase should be listed in one column, followed by independent model building from the second phase, followed by hospital validation for the first phase, then hospital validation for the second phase. As currently formatted, the Table appears to have aggregated both phase 1 and 2 in temporal and both phases along with hospital validation in the second instance. Also, according to Table S3, the 2008 data comprised only 0.1% of all data, so it is really a reflection of 2007 (about 25% of all data). Need to compare rates of missing data for phase 1 and phase 2 for all and for the key variables of interest in Table format.

5. Fig 2: The Authors included ROC curves, but chose to omit sensitivity and specificity (page 9, last para), upon which ROC depends. If these are to be included as part of results and discussion, then need to include a Table (in main text) of the 4 models ROC, sens, spec,
all with CIs.

6. Fig 3: The calibration curves are incomplete. This particular figure could be placed in supplemental material. A more important calibration curve would compare a statistical model with a machine learning model on the same grid, but with the format suggested by TRIPOD "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and Elaboration" by K.G.M. Moons, D.G. Altman, J.B. Reitsma, et al, Annals of Internal Medicine 2015:162:W1-W73.. In particular, the Fig 8 of that article should be emulated. That figure allows the reader to see the relationship of observed vs predicted probabilities along the spectrum of probabilities from the data, along with confidence intervals for those prediction estimates. The advantage to this level of detail is that it would convey to the reader the strength of association at various model scores, along with their relative uncertainty, reflecting how many data were available at various cut-points. Surely, the model was based on more information at the lower probabilities and less at the higher probabilities, thus the CIs are wider at the higher end. A line or curve alone does not convey that information, nor does the C-statistic alone. The reader should be able to visually compare the model performance along the spectrum of prediction probabilities.

Associate Editor's Comments:

Please in your revision be particularly responsive to Reviewer #4 (the Statistical Editor).

EDITORIAL OFFICE COMMENTS:

1. The Editors of Obstetrics & Gynecology are seeking to increase transparency around its peer-review process, in line with efforts to do so in international biomedical peer review publishing. If your article is accepted, we will be posting this revision letter as supplemental digital content to the published article online. Additionally, unless you choose to opt out, we will also be including your point-by-point response to the revision letter. If you opt out of including your response, only the revision letter will be posted. Please reply to this letter with one of two responses:
A. OPT-IN: Yes, please publish my point-by-point response letter.
B. OPT-OUT: No, please do not publish my point-by-point response letter.

2. As of December 17, 2018, Obstetrics & Gynecology has implemented an "electronic Copyright Transfer Agreement" (eCTA) and will no longer be collecting author agreement forms. When you are ready to revise your manuscript, you will

be prompted in Editorial Manager (EM) to click on "Revise Submission." Doing so will launch the resubmission process, and you will be walked through the various questions that comprise the eCTA. Each of your coauthors will receive an email from the system requesting that they review and electronically sign the eCTA.

Please check with your coauthors to confirm that the disclosures listed in their eCTA forms are correctly disclosed on the manuscript's title page.

3. In order for an administrative database study to be considered for publication in Obstetrics & Gynecology, the database used must be shown to be reliable and validated. In your response, please tell us who entered the data and how the accuracy of the database was validated. This same information should be included in the Materials and Methods section of the manuscript.

4. Responsible reporting of research studies, which includes a complete, transparent, accurate and timely account of what was done and what was found during a research study, is an integral part of good research and publication practice and not an optional extra. Obstetrics & Gynecology supports initiatives aimed at improving the reporting of health research, and we ask authors to follow specific guidelines for reporting randomized controlled trials (ie, CONSORT), observational studies (ie, STROBE), meta-analyses and systematic reviews of randomized controlled trials (ie, PRISMA), harms in systematic reviews (ie, PRISMA for harms), studies of diagnostic accuracy (ie, STARD), meta-analyses and systematic reviews of observational studies (ie, MOOSE), economic evaluations of health interventions (ie, CHEERS), quality improvement in health care studies (ie, SQUIRE 2.0), and studies reporting results of Internet e-surveys (CHERRIES). Include the appropriate checklist for your manuscript type upon submission. Please write or insert the page numbers where each item appears in the margin of the checklist. Further information and links to the checklists are available at http://ong.editorialmanager.com. In your cover letter, be sure to indicate that you have followed the CONSORT, MOOSE, PRISMA, PRISMA for harms, STARD, STROBE, CHEERS, SQUIRE 2.0, or CHERRIES guidelines, as appropriate.

5. Standard obstetric and gynecology data definitions have been developed through the reVITALize initiative, which was convened by the American College of Obstetricians and Gynecologists and the members of the Women's Health Registry Alliance. Obstetrics & Gynecology has adopted the use of the reVITALize definitions. Please access the obstetric and gynecology data definitions at https://www.acog.org/About-ACOG/ACOG-Departments/Patient-Safety-and-Quality-Improvement/reVITALize. If use of the reVITALize definitions is problematic, please discuss this in your point-by-point response to this letter.

6. Because of space limitations, it is important that your revised manuscript adhere to the following length restrictions by manuscript type: Original Research reports should not exceed 22 typed, double-spaced pages (5,500 words). Stated page limits include all numbered pages in a manuscript (i.e., title page, précis, abstract, text, references, tables, boxes, figure legends, and print appendixes) but exclude references.

7. Specific rules govern the use of acknowledgments in the journal. Please note the following guidelines:

* All financial support of the study must be acknowledged.
* Any and all manuscript preparation assistance, including but not limited to topic development, data collection, analysis, writing, or editorial assistance, must be disclosed in the acknowledgments. Such acknowledgments must identify the entities that provided and paid for this assistance, whether directly or indirectly.
* All persons who contributed to the work reported in the manuscript, but not sufficiently to be authors, must be acknowledged. Written permission must be obtained from all individuals named in the acknowledgments, as readers may infer their endorsement of the data and conclusions. Please note that your response in the journal's electronic author form verifies that permission has been obtained from all named persons.
* If all or part of the paper was presented at the Annual Clinical and Scientific Meeting of the American College of Obstetricians and Gynecologists or at any other organizational meeting, that presentation should be noted (include the exact dates and location of the meeting).

8. Provide a précis on the second page, for use in the Table of Contents. The précis is a single sentence of no more than 25 words that states the conclusion(s) of the report (ie, the bottom line). The précis should be similar to the abstract's conclusion. Do not use commercial names, abbreviations, or acronyms in the précis. Please avoid phrases like "This paper presents" or "This case presents."

9. The most common deficiency in revised manuscripts involves the abstract. Be sure there are no inconsistencies between the Abstract and the manuscript, and that the Abstract has a clear conclusion statement based on the results found in the paper. Make sure that the abstract does not contain information that does not appear in the body text. If you submit a revision, please check the abstract carefully.

In addition, the abstract length should follow journal guidelines. The word limits for different article types are as follows: Original Research articles, 300 words. Please provide a word count.

10. Only standard abbreviations and acronyms are allowed. A selected list is available online at http://edmgr.ovid.com/ong/accounts/abbreviations.pdf. Abbreviations and acronyms cannot be used in the title or précis. Abbreviations and acronyms must be spelled out the first time they are used in the abstract and again in the body of the manuscript.

11. The journal does not use the virgule symbol (/) in sentences with words. Please rephrase your text to avoid using "and/or," or similar constructions throughout the text. You may retain this symbol if you are using it to express data or a measurement.

12. In your Abstract, manuscript Results sections, and tables, the preferred citation should be in terms of an effect size, such as odds ratio or relative risk or the mean difference of a variable between two groups, expressed with appropriate confidence intervals. When such syntax is used, the P value has only secondary importance and often can be omitted or noted as footnotes in a Table format. Putting the results in the form of an effect size makes the result of the statistical test more clinically relevant and gives better context than citing P values alone.

If appropriate, please include number needed to treat for benefits (NNTb) or harm (NNTh). When comparing two procedures, please express the outcome of the comparison in U.S. dollar amounts.

Please standardize the presentation of your data throughout the manuscript submission. For P values, do not exceed three decimal places (for example, "P = .001"). For percentages, do not exceed one decimal place (for example, 11.1%").

13. Please review the journal's Table Checklist to make sure that your tables conform to journal style. The Table Checklist is available online here: http://edmgr.ovid.com/ong/accounts/table_checklist.pdf.

14. The American College of Obstetricians and Gynecologists' (ACOG) documents are frequently updated. These documents may be withdrawn and replaced with newer, revised versions. If you cite ACOG documents in your manuscript, be sure the reference you are citing is still current and available. If the reference you are citing has been updated (ie, replaced by a newer version), please ensure that the new version supports whatever statement you are making in your manuscript and then update your reference list accordingly (exceptions could include manuscripts that address items of historical interest). If the reference you are citing has been withdrawn with no clear replacement, please contact the editorial office for assistance (obgyn@greenjournal.org). In most cases, if an ACOG document has been withdrawn, it should not be referenced in your manuscript (exceptions could include manuscripts that address items of historical interest). All ACOG documents (eg, Committee Opinions and Practice Bulletins) may be found via the Clinical Guidance & Publications page at https://www.acog.org/Clinical-Guidance-and-Publications/Search-Clinical-Guidance.

15. Authors whose manuscripts have been accepted for publication have the option to pay an article processing charge and publish open access. With this choice, articles are made freely available online immediately upon publication. An information sheet is available at http://links.lww.com/LWW-ES/A48. The cost for publishing an article as open access can be found at http://edmgr.ovid.com/acd/accounts/ifauth.htm.

Please note that if your article is accepted, you will receive an email from the editorial office asking you to choose a publication route (traditional or open access). Please keep an eye out for that future email and be sure to respond to it promptly.

***

If you choose to revise your manuscript, please submit your revision through Editorial Manager at http://ong.editorialmanager.com. Your manuscript should be uploaded in a word processing format such as Microsoft Word. Your revision's cover letter should include the following:

* A confirmation that you have read the Instructions for Authors (http://edmgr.ovid.com/ong/accounts/authors.pdf), and
* A point-by-point response to each of the received comments in this letter.

If you submit a revision, we will assume that it has been developed in consultation with your co-authors and that each author has given approval to the final form of the revision.

Again, your paper will be maintained in active status for 21 days from the date of this letter. If we have not heard from you by Jan 10, 2020, we will assume you wish to withdraw the manuscript from further consideration.

Sincerely,

The Editors of Obstetrics & Gynecology

2018 IMPACT FACTOR: 4.965
2018 IMPACT FACTOR RANKING: 7th out of 83 ob/gyn journals

_____

We thank the editors and reviewers for their favorable review of our manuscript. Below we respond to each query in red and bold font, and all changes in the text of the manuscript have been made in tracked changes.

REVIEWER COMMENTS:

Reviewer #1:

1. Overall this paper addresses a timely and important issue in obstetrics through application of novel methods. The main challenge the authors of this paper face is the task of communicating complex statistical and machine learning methods to the average practicing Ob/Gyn in an approachable and informative way. Towards this end, the authors spend a considerable portion of the paper explaining their methods and teaching the reader about how these models are developed, interpreted and evaluated. Overall, the authors accomplish this task, though I think a few points could be further clarified which I will delineate below.

Thank you for this favorable review. As requested by the statistical reviewer below, we have also added references to two recent manuscripts that describe how to read and interpret manuscripts that use machine learning methods for the clinician who may be unfamiliar with these techniques (Bi, American Journal of Epidemiology, 2019 and Liu, JAMA, 2019). We have also added this text to the methods to highlight these resources for readers when interpreting the current study (Line 174): "Recent review articles provide a framework for interpreting clinical studies that use machine learning methods for clinical readers."

2. Introduction: For the background portion of the introduction, I would consider adding in a short statement about how accurate prediction of adverse events such as PPH can allow for improved clinical outcomes. Additionally, the authors mention rates of transfusion though do explicitly state the utility of their models for predicting transfusion rates.

We have added the following statement to the introduction (Line 123): "and possibly improved clinical outcomes." Whether accurate prediction will impact clinical outcomes will need to be studied and confirmed. In the results, we provide the frequency of transfusion among women with vs. without hemorrhage. In the discussion, we have added the following (Line 347): "Our definition of hemorrhage did not incorporate transfusion of blood products and further models to accurately predict transfusion beyond current guidelines are needed as recently highlighted (Kawakita, OBGYN, 2019)."

3. The third paragraph of the introduction contrasting traditional statistical models and machine learning is a bit of an abrupt transition. A statement such as "Current methods for predicting PPH are based on risk stratification methods. However improved predictive ability could be achieved through application of traditional statistical models and machine learning. Recent advances in computer science have led to the development of machine learning…which has superior predictive ability over traditional statistical

methods." It is important to the paper to situate these two approaches (statistical and machine learning models) in relation to current, less individualized risk stratification approaches. Excellent explanation of machine learning!

We have rephrased this paragraph, as suggested (Line 111): "Current methods for predicting postpartum hemorrhage are based on risk stratification methods. Improved predictive ability could be achieved by applying traditional statistical and machine learning methods."

4. Methods: Consider changing the description of the CSL database to present tense (vs past tense).

   We have kept the description in past tense so that it is consistent throughout the methods given we use past tense throughout. There are some instances where we have now changed the tense from present to past in the methods to be consistent (see tracked changes).

5. For the first paragraph on page 9, consider clarifying that logistic regression is the first of the two statistical models. The remainder of this paragraph includes a very brief explanation of the two machine learning models and includes several terms that would be unfamiliar to the average reader. I would consider either further explaining these terms ("bootstrap samples", "training data," "ensembles") or perhaps leaving these explanations out entirely and making a more general statement that two types of machine learning models were created as it is not entirely necessary that the reader understand the models—only the conclusion that such models can be made and have good predictive ability.

   Yes, we clarify that logistic regression is the first of the two statistical models. These terms are described further in the recent reviews of machine learning for clinical readers we now cite early in the methods (Line 174): "Recent review articles provide a framework for interpreting clinical studies that use machine learning methods for clinical readers." We have also further condensed the descriptions of the machine learning methods as suggested. Please see tracked changes lines 163 to 166.

6. In the last paragraph on page 9 (begins "Measures, such as sensitivity…") explaining how model performance is assessed the authors do an excellent job of educating the reader and positioning them to understand what follows.

   Thank you for this favorable comment.

7. The last paragraph on page 12 (begins "Figure 3 demonstrates…") and the first paragraph on page 13 (begins "In the decision curve analysis…") it is unclear to me what the following means: "This model assigned an accurate probability of post partum hemorrhage when prediction ranged from 0 to 70-80%." Perhaps it is clear to those more familiar with these types of tests/statistics.

We have now changed the paragraph in line with the comments of the statistical editor below requesting calibration curves for all four models, and also describe the findings further (Line 273): "Figure 2 demonstrates the overall calibration curves with 95% confidence intervals of the four models. The calibration curve shows the variation in performance of each model in comparison to perfect agreement between the predicted probability of the model and the actual probability. We also present the best performing model, Extreme Gradient Boosting, by each of the ten assessed sites in Appendix Figure 2."

8. For the last paragraph of the results section on page 13 (begins "We were unable to compare our models…") I would consider additional clarification as this paragraph addresses the secondary aim of the study. A suggestion "We were unable to compare our models to current risk stratification strategies as outlined by CMQCC (i.e., low, medium, and high risk for post partum hemorrhage) as only two thirds of the variables included in CMQCC guidelines were identified by our models and many of the variables included in CMQCC (such as uterine fibroids, known bleeding disorder or coagulopathy) were not routinely assessed on labor admission and therefore were not considered as candidate variables for our models."

   We have added the additional statement as suggested above (Line 294) "and therefore were not considered as candidate variables for our models."

9. Discussion: While the authors clearly state that machine learning models outperformed the statistical models, I would consider including a more robust discussion of what level of discrimination is clinically relevant. This is particularly salient as they immediately state that as statistical models are more easily incorporated into practice in EMRs that perhaps statistical models, rather than machine learning models, would be a reasonable alternative as they performed reasonably well in predictive ability.

   We have added the following with regards to the level of discrimination that is clinically relevant (Line 331): "and a final model should be chosen based on a combination of discrimination, calibration (especially accurate calibration in the range of predictions where clinical decisions will be affected), ease of use in the clinical setting, and acceptability by clinicians and patients."

10. The authors make a general statement that "any model will need to be prospectively assessed in local contemporary cohorts of pregnant women across regional settings." Do the authors intend to further prospectively validate these models?

    We plan to apply the models built here using available EHR data within our institution and encourage others to do so. We have also provided a supplemental table with the variables used in each of the four models as supplementary table S4 cited in line 260.

Reviewer #2:

1. The authors aim to construct models to predict PPH using both machine learning and more traditional statistical modeling. I have the following comments regarding the manuscript.

   Thank you for this variable review. We address each comment below.

2. Abstract: The Conclusion about these models allowing providers to transfer women to appropriate levels of maternal care goes way beyond what can be concluded from this particular project.

   We have replaced this sentence so as to not overstate the conclusion (Line 87): "Further clinical application is needed, which may assist providers to be prepared and triage at risk women to the appropriate level of maternity care."

3. Intro: Consider deleting the possible secondary aim of comparison to the CMQCC risk stratification. We all brainstorm research ideas and then go about determining if they are feasible. Since this was found not to be feasible, it does not need to be included in the manuscript other than perhaps a mention in the discussion of the limitations of the CSL dataset and inability to compare to other risk stratification systems.

   Yes, as requested we have removed this secondary aim from the introduction (Line 127). The results of this analysis are stated in the results section for interested readers.

4. Methods: The authors constructed two machine learning models, and two more traditional statistical models. Was one of the objectives to compare these different methods for predicting PPH?

   Yes models were also compared, and we have added (Line 63) "compare" to the first sentence of the methods.

5. Can the authors clarify how lasso regression provides a "different method" of identifying the final variables in the model compared to standard logistic regression?

   We have added (Line 165): "Lasso regression is also referred to as penalized regression because a penalty is imposed on variables with high variance in order to eliminate the number of variables and improve model predictions."

6. The paragraph about the decision not to use sensitivity, specificity, etc could be significantly pared down. Just tell the reader what you did to analyze the data.

   We have decided to keep his content given the comment below from the statistics editor about not presenting this information. We have moved this paragraph to line 212 so that it does not detract from the prior description of methods.

7. Consider first saying that decision curve analysis (DCA) was used to quantify the net benefit of the models, and then further describe DCA. Otherwise the authors are well into the description of DCA before the reader knows what it was used for in this manuscript.

   Yes, as suggested we have moved the ordering of sentences as suggested (Line 200).

8. How did the authors deal with missing data?

   We state in the methods (Line 210): "Missing predictor values were imputed using Multiple Imputed Chained Equations (MICE)." We have also added an additional analysis assessing missing data over time as requested by the statistical editor (line 218): "The frequency of missing data for each variable generally did not vary over time between the two time periods used in temporal validation (2002-2006 and 2007-2008) (Appendix Table 4)."

9. Results: A 2% rate of blood transfusion for births WITHOUT PPH seems quite high. Do the investigators suspect a data collection error? Please explain.

   We have added to the discussion (Line 349): "We did note a 2% rate of transfusion among those women without postpartum hemorrhage, which may reflect outcome misclassification or cases of transfusion that occurred after delayed or late postpartum hemorrhage."

10. The authors compare the models based on their discriminatory capability and report higher or lower discriminatory capacity. Can the c-statistic be compared statistically?

    There are statistical tests to compare c-statistics, such as the DeLong test to compare the change in AUC. However, we decided not to use this approach given limitations of this approach when comparing two nested models (i.e., observe that adding new risk factors to a model that contains 'standard' predictors results in two 'nested' models), which can lead to loss of power (Demler et al, *Statistics in Medicine*, 2012).

11. Would recommend being very precise about when you are using individual hospitals versus sites. For example text about figure 3 says you are comparing sites and the legend says hospitals. Would double check this throughout.

    Yes, this is an important clarification, and in our analyses, we have compared sites, not hospitals, of which multiple hospitals could be within a site (the dataset included 19 hospitals within 12 sites). Hence, we have replaced the term hospital with site throughout the paper as appropriate.

12. Why is the upper limit of prediction listed at 70-80%? Is it because prediction was broken into deciles for analysis? If so, this should be reported in the methods.

This figure has now been redone to show the calibration curves for all for models ranging from a predicted probability of 0.0 to 1.0.

13. The idea of net benefit is not presented comprehensively until the legend of Figure 4. Why aren't these equations in the methods?

    We have added the following text to the methods, consistent with the legend from Figure 3 (Line 211): "The net benefit is calculated as: Net benefit = true positive rate - (false positive rate x weighting factor) in which the weighting factor = Threshold probability/1-threshold probability and the threshold probability is a level of certainty above which the patient or physician would choose to intervene."

14. Was the prenatal diagnosis of macrosomia in CSL by Leopold's or ultrasound? Was it universally assessed? I doubt it since the CSL data were pulled retrospectively.

    This is a limitation of this large dataset, which will likely also be a limitation when this model is applied to EHR data. We have added the following limitation to the discussion (Line 381): "Some assessed variables, such as a prenatal diagnosis of macrosomia, were likely not universally captured and it was unclear how they were defined."

15. Figure 3. Have the authors looked in detail at data from hospital 41 to ensure that it does not appear to be flawed/missing, etc? It is clearly out of line when compared to other hospitals.

    Yes, we assessed data from site 41 and confirmed that data was not flawed nor missing. This figure has now been moved to the appendix as Appendix Figure 1 as suggested by the statistical editor to highlight variation by site in our results.

16. Figure 5. Why abbreviate the predictors on the y-axis with non-intuitive abbreviations

    The abbreviations have now been removed. Please refer to the updated figure.

17. Discussion: Last paragraph is appropriately tempered to convey potential down the road uses. Need to tone down the idea of transfer to higher level of care based on these models.

    Yes, we have further tempered this statement, including in the conclusion of abstract.

18. 50 references is a lot for original research. Can the list be shortened to critical references?

    The number of references is longer for this manuscript as we are citing not only relevant studies with regards to postpartum hemorrhage and clinical prediction of obstetrical outcomes, but are also summarizing machine learning methods and

terminology for a clinical readership who are likely unfamiliar with these newer statistical techniques.

19. If the actual models are not provided, how can the reader use them or validate them?

These models are available to readers upon request. We have added an additional supplemental Table S5 where we provide the variables included in each of the four models (Line 259): "The variables included in each model are presented in Appendix Table 5." We also provide the most important variables in the machine learning model in Figure 4. We anticipate that these models can be made available in the future via an appropriately supported electronic public platform. As we state in the discussion (line 306): "It would be reasonable to integrate the models into an online calculator or automated input in the electronic medical record for immediate use on labor admission."

Reviewer #3:

1. In this manuscript, the authors describe the application of a relatively new approach to analyze data regarding the common obstetric complication of postpartum hemorrhage (PPH). They use a large, publicly available dataset to build and compare models for prediction of individual patient PPH risk based on information available at the time of admission. Although this has been done previously with other datasets, the major novelty of this project is the use of two different machine learning approaches, in addition to statistical models, and compared their accuracy for prediction in their validation process. Indeed, one of the major strengths of this manuscript is that it could be one of the first publications in our field to demonstrate the potential for machine learning to create highly accurate predictive models for important individual patient outcomes. This would presumably provide superior performance to current guidelines which are based in some part on expert opinion without validation. Another strength is the considerations elaborated in the Discussion regarding limitations and further developments needed before an optimal model is ready for clinical use.

   Thank you for the favorable review of our manuscript.

2. The methods section contains a very complex, extended description of what the models entailed and the rationale for the use of four different models. The descriptions of how the analyses should be interpreted is very much appreciated. Full understanding of the methods will be beyond most readers. Yet, it can be appreciated that data supports the conclusions reached.

   As also requested by the statistical reviewer, we have also added references to two recent manuscripts that describe how to read and interpret manuscripts that use machine learning methods for the clinician or perinatal epidemiologist who may be unfamiliar with these techniques (Bi, American Journal of Epidemiology, 2019 and Liu, JAMA, 2019). We have also added this text to the methods to highlight these resources for readers when interpreting the current study (Line 174): "Recent

publications provide a framework for interpreting clinical studies that use machine learning methods for clinical readers."

3. The manuscript does specifically state that the prediction is based on factors present at the time of patient admission. There are recognized intrapartum risk factors that can affect the risk of PPH, such as prolonged labor/stimulation with oxytocin and intrapartum chorioamnionitis, which are therefore not included. With the predictive accuracy described, it is hard to imagine that the C-statistic could get much higher; but presumably a model which could incorporate important developments after admission would have more clinical value at the time of delivery or after. And as stated in the Introduction, to objective was to develop and validate a predictive model for PPH, not limited to those available at the time of admission. An argument can be made for performing the analysis upon admission (for the purpose of This could be discussed as part of limitations.

   The aim of the current analysis was to develop a model that could be used on labor and delivery admission, and hence all variable that were assessed were those that would be available on admission (Line 125). In the introduction, for the primary aim we have added "on labor admission" (Line 125). In the discussion, we have added (Line 313): "Our model focused on prediction at the time of labor admission and hence employed variables available at that time, and it may also be reasonable to build models that include intrapartum variables that affect hemorrhage risk, such as length of labor and mode of delivery."

4. Methods—it is not clear from the text that the elements listed in Appendix Table 3 are the elements used in the modeling. This table is difficult to read because of the way it is formatted. Consider indent every other line with the Missing n so that the actual elements appear lined up to the left margin.

   We have added the following to the text in the results "for possible model inclusion" with regards to the variables in Appendix Table 3 (Line 250). We have also added (Line 260): "The variables included in each model are presented in Appendix Table 5." To ease with reading this table, as requested, we have left indented each missing n label.

5. Results—in the second paragraph of the results, 14 variables are listed in the first sentence. The second sentence contains another 16 or so variables are listed and described as associated with PPH. Presumably we are to infer that those listed in the first sentence are not associated with PPH. It is not clear why you list this many of them, but not all 55.

   The first sentence has now been simplified and we no longer list 14 variables (Line 249): "A total of 55 candidate predictors of postpartum hemorrhage available on labor admission were assessed for possible model inclusion, including socio-demographic, obstetric, clinical, and physiologic variables (Appendix Table 3)."

6. Results—in the fourth paragraph, it would be helpful if it were explained that this analysis refers to the data presented in the second column of table 1.

   As requested by the statistical editor, this table has now been redone to show temporal validation and then temporal and site validation separately. We have added in this paragraph (Line 262): "After temporal and site validation, . . ." so it is clear we are referring to results from the second set of columns.

7. Figure 3—the legend is very helpful in understanding what is being done, but more explanation in the text would be helpful as to how the accuracy probably was assigned, and what the meaning of this is.

   We have changed the text as follows (Line 274): "Figure 2 demonstrates the overall calibration curves with 95% confidence intervals of the four models. The calibration curve shows the variation in performance of each model in comparison to perfect agreement between the predicted probability of the model and the actual probability. We also present the best performing model, Extreme Gradient Boosting, by each of the ten assessed sites in Appendix Figure 2."

8. Figure 4—similarly, please describe in the text what is seen in the figure which allows for the conclusion being made.

   As requested by a previous reviewer, we have added the following text to the methods which is adapted from the Figure 3 legend (Line: 212): "The net benefit is calculated as: Net benefit = true positive rate - (false positive rate x weighting factor) in which the weighting factor = Threshold probability/1-threshold probability and the threshold probability is a level of certainty above which the patient or physician would choose to intervene."

STATISTICAL EDITOR'S COMMENTS:

1. Generalizability: As can be seen in Table S3 (some of this information should be shown more prominently in main text), those with missing data for EBL (and therefore omitted from the model analyses) were ~ 1/3 of the original cohort, with 67.0% available for vaginal and 65.8% for cesarean deliveries. This proportion of missing data is a major limitation to general application of these conclusions. Specifically for the model building, some important variables had high rates of missing data (eg, pre-pregnancy wgt (29.4% missing overall, with 45.8% missing among the PPH subset). Proportions less, but still important for another important variable, adm wgt: 15.7% missing overall with 23.0% missing among the PPH cohort. Also, for macrosomia, the overall cohort had 44.7% missing data, while for PPH, the missing data rate was 5.6%. These three variables were cited in Fig 5 as comprising the highest importance values for the nominally best prediction model for PPH. Even if these were imputed (pg 10, last line, ref (35)), the proportion of missing data seriously limits imputation and therefore how this model faithfully reflects the population studied.

Yes, this is an important limitation which we have added to the discussion (Line 354): "Missing data is an important limitation of this analysis, including restricting this analysis to the subset of the original cohort with available blood loss data and the substantial proportion of covariates with incomplete data, albeit we used up-to-date imputation techniques. The proportion of missing data is a limitation to the general application of these models affecting generalizability of these results. However, it is likely that missing data or incomplete ascertainment will continue to be a limitation when applying these models in real time with EHR data."

2. General: Machine learning would be new topic for most readers, so there should be a short primer on its methods (could be supplemental material) with concise summary in main text.

As also requested by the reviewers #1 and #2, we have also added references to two recent manuscripts that describe how to read and interpret manuscripts that use machine learning methods for the clinician or perinatal epidemiologist who may be unfamiliar with these techniques (Bi, American Journal of Epidemiology, 2019 and Liu, JAMA, 2019). We have also added this text to the methods to highlight these resources for readers when interpreting the current study: "Recent publications provide a framework for interpreting clinical studies that use machine learning methods for clinical readers."

3.It is unclear from Methods whether the same variables were retained in each of the 4 models evaluated. If so, should so state. If not, then should construct a Table listing the retained variables in each model in rank order, with the relative importance of each to the final prediction. Also, it seems from the general results that cesarean delivery is associated with higher risk of PPH than vaginal delivery. Why was CD not included as a potential variable, or analyzed as a separate cohort, since > 90% of the primary outcome was associated with CD? That variable alone would have high discriminatory value for prediction of PPH.

We have provided an additional supplemental table, now labeled S5, where we list the variables used in each of the 4 models. In Figure 4, we provide the listing of the most important variables in rank order for Extreme Gradient Boosting, the best performing model. The focus of our analysis was prediction at the time of labor admission at which time mode of delivery would generally have not been decided, and hence intrapartum factors, including mode of delivery, were not included. However, it may be reasonable to also build prediction models that include intrapartum variables in which case mode of delivery could be considered. We have added the following to the discussion (Line 312): "Our model focused on prediction at the time of labor admission and hence employed variables available at that time, and it may also be reasonable to build models that include intrapartum variables that affect hemorrhage risk, such as length of labor and mode of delivery."

4. Table 1: This Table needs much more elaboration. The model results (C-statistic with CIs) from the first phase should be listed in one column, followed by independent model building from the second phase, followed by hospital validation for the first phase, then hospital

validation for the second phase. As currently formatted, the Table appears to have aggregated both phase 1 and 2 in temporal and both phases along with hospital validation in the second instance. Also, according to Table S3, the 2008 data comprised only 0.1% of all data, so it is really a reflection of 2007 (about 25% of all data). Need to compare rates of missing data for phase 1 and phase 2 for all and for the key variables of interest in Table format.

This table has now been redone as suggested with temporal validation and then temporal and site validation shown separately. Given the concern for missing data, we have also added an additional appendix table S4 where we present the proportion of missing data for each variable during the two time periods (Line 219): "The frequency of missing data for each variable generally did not vary over time between the two time periods used in temporal validation (2002-2006 and 2007-2008) (Appendix Table 4)."

5. Fig 2: The Authors included ROC curves, but chose to omit sensitivity and specificity (page 9, last para), upon which ROC depends. If these are to be included as part of results and discussion, then need to include a Table (in main text) of the 4 models ROC, sens, spec, all with CIs.

We have further highlighted in the methods why we have chosen not to present sensitivity and specificity results (Line 222): "Measures, such as sensitivity, specificity, and false positive and negative probabilities are not recommended when reporting performance of clinical prediction models because they are performance measures after introducing one or more probability thresholds. While these are useful for estimating accuracy or classification measures often reported in a single diagnostic test or prognostic factor studies, such dichotomization and related classification measures lead to loss of information when providing a prediction for the future and introducing such a threshold implies that it is relevant to clinical practice, which often is not the case."

For this reason, we have also removed the ROC curve Figure from the manuscript and placed it in the appendix as appendix figure 1.

6. Fig 3: The calibration curves are incomplete. This particular figure could be placed in supplemental material. A more important calibration curve would compare a statistical model with a machine learning model on the same grid, but with the format suggested by TRIPOD "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and Elaboration" by K.G.M. Moons, D.G. Altman, J.B. Reitsma, et al, Annals of Internal Medicine 2015:162:W1-W73.. In particular, the Fig 8 of that article should be emulated. That figure allows the reader to see the relationship of observed vs predicted probabilities along the spectrum of probabilities from the data, along with confidence intervals for those prediction estimates. The advantage to this level of detail is that it would convey to the reader the strength of association at various model scores, along with their relative uncertainty, reflecting how many data were available at various cut-points. Surely, the model was based on more information at the lower probabilities and less at the higher probabilities, thus the CIs are wider at the higher end. A line or curve alone does not convey that information, nor does the C-statistic alone. The

reader should be able to visually compare the model performance along the spectrum of prediction probabilities.

We have revised Figure 3 to be consistent with the statistical editor's recommendation of showing curves from the statistical models and machine learning models on the same image as suggested by the TRIPOD guidelines in order to compare the four models. Please refer to the revised Figure 2. The current figure 2 has now been placed in supplemental material as appendix figure S2.

Associate Editor's Comments:

Please in your revision be particularly responsive to Reviewer #4 (the Statistical Editor).

Yes, we have aimed to address the comments of all four reviewers, including the Statistical Editor.