PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The predictive value of machine learning on fracture risk in osteoporosis:a systematic review and meta-analysis
AUTHORS	Wu, Yanqian; Chao, Jianqian; Bao, Min; Zhang, Na

VERSION 1 – REVIEW

REVIEWER	Carey, John Galway University Hospitals, Rheumatology
REVIEW RETURNED	19-Jan-2023

GENERAL COMMENTS	Thank you for asking me to review this article reviewing the predictive value of machine learning on fracture risk in osteoporosis. The authors performed a substantial body of work, some of which is presented well. However parts are also confusing, and in my opinion this article would benefit from significant modification before being considered. It could be very interesting if it were shorter, clearer and focussed on the message in the title. I have some major concerns about the authors understanding of the data and terminology used resulting in a lack of clarity and consistency for robust interpretation. Major Concerns: Machine learning is a “method” or “technique”, not a “model” and this needs to be corrected and clarified throughout the paper. The model(s) generated will depend on the data, the characteristics of the study population and outcome of interest, the methods and analyses performed. Applying the same ‘technique’ to different datasets will inevitably lead to the production of different results; if the results are robust however the differences should be smaller, and results more consistent. This needs due consideration here. In addition if the same ‘technique’ is applied to different datasets, it may result in different models being generated, particularly when there are different variables in the different datasets. This key aspect receives little attention and should be addressed as it will help explain some of the heterogeneity between studies. The c-statistic has important limitations, particularly with respect to risk rather than discrimination, so this needs some discussion. This, and the over simplification using a Forest Plot in figure 3 needs further discussion. All these results are very dependent on multiple factors as outlined above, and also sample size, generalisability and frequency and accuracy of outcome/events. The results as presented appear overly optimistic, as some studies have produced much lower values, e.g. Kong et al: PMID: 32161842. Many clinicians are less familiar with artificial intelligence, machine learning and deep learning and a little more information on what

	machine learning actually is, how it differs from standard techniques, and the differences between supervised and unsupervised ML and why they matter is important for a journal such as this. Greater clarity is needed around this, for example a summary of how these various methods differ, perhaps a table? e.g. 'handles missing data' etc. Their work shows published results are markedly heterogeneous in rigour, design, technique and application and "Algorithmovigilance" is urgently needed (PMID: 33856479). I am concerned the first sentence in the discussion section is misleading as this paper doesn't appear to include all currently available fracture risk prediction methods or models. In addition unless there are papers directly comparing different techniques and resulting model performance on the same datasets we cannot deduce which technique performs best. It would be better to consider studies which used a single method separately to those which compared different techniques in order to provide substance to this deduction such as the study by Kruse et al (reference 39) and then provide some sort of ranking, if of course that is possible. It may well be that different techniques perform very differently in different populations or on different datasets due to some of the reasons outlined in the discussion such as handling of missing data, smaller numbers, over or under representation of particular variables or outcomes and skewed/non-parametric predictors. The authors allude to some of these shortcomings in their discussion, but this could be more coherent and precise. In paragraph 7 of the discussion section the authors state 'existing ML models for fracture prediction focus on populations in Western countries' in fact of the studies they included 15 are European, 17 North American and 14 Asia-Pacific. I don't agree the data support their conclusion as their results suggest the field is very heterogeneous and sorely in need of a better structure and process. Surely the benefit of ML would be to accelerate the development of better models for specific populations rather than a 'one size fits all'? Minor Concerns: There are many, so I only include some more salient ones here. The authors could do with some independent help to review their grammar and terminology for this publication. Figure 1. In box 3 there were 1673 records, n = 0 not retrieved, and in the next box (4) there are only 340 reports assessed! What happened to the other 1,433?? (No mention in the results narrative either). The second paragraph of the "Results" needs to be rewritten / tidied up. The author state there were forty-six studies. Stick to numbers or words for consistency. Then second sentence of this paragraph: "The majority of the studies were conducted in U.S. (n=10) and Canada (n=6)," needs to be reworded as this only represents 35% of the studies; ?most common location. Next part of this sentence: "most were cohort studies (n = 40) or case-control studies (n = 6)" should read: "most were cohort studies (n = 40), and the rest were case-control studies (n = 6)"? "Most study samples covered postmenopausal women (n=15)" is incorrect as this represents only 33% of the studies? Do you mean the number of subjects or the number of studies? In my opinion it would be nice to include the sample size in table 1. This would be preferable to the author name as each study can be retrieved once referenced. This is included in Table S3 but I think

	Table 1 should include this, and look neater and tidier like tables S3 and S4. Consider deleting the word “data” from the rows in the column titled “Data Source” and put in things like “Database”, “Registry”, “Medical Records” etc. The wording used in the next column is very confusing under “Sample Population Type”. I am unclear what the difference between a “patient”, “older women”, “postmenopausal women”, “older men”, “inhabitant”, “elderly” etc. is. Usual terms include: “men”, “women”, “patients”, “subjects”. Please simplify and clarify. Consider substituting “mean” for “average” for clarity in the Age column title. In the “Fracture site” Column would it not be better to just list as “hip”, “vertebral”, “multiple” rather than repeatedly including the word “fracture”? The “ML Models” are not really “models” but “methods”. The final model will be derived using these various machine learning methods, e.g. LR, ANN etc. Very little data is presented on the ‘models’. Paragraph 4: “We roughly classify”. Please delete “roughly” and clarify. I am unclear why fracture history and falls are included with “demographics”, while “osteoporosis” and “fracture type” are included under “Comorbidity”. This needs tidying up. Recommend keeping demographics to demographics, and clinical risk factors and comorbidities together. Please clarify if bone mineral density is in ‘grams/centimeter2’ or the ‘T-score’, or both. Figure 2 appears incorrect: 1. “Low” is not included in the “overall” bar (5th column/bar) 2. There is a yellow square box in the figure legend with no wording beside it, Omit?? While I like the concept of a Forest Plot to summarise lots of data, I am concerned figure 3 is difficult to understand/interpret. The authors evaluating ML ‘methods’ to develop ‘models’ but provide no information on what was actually in these models, or how they compare to any other model, e.g. FRAX. A plot like this may have some value if it only included results from ‘validation’ datasets, though some of the same concerns remain.

REVIEWER	Vergani, Laura Maria Politecnico di Milano, Mechanical Engineering
REVIEW RETURNED	23-Jan-2023

GENERAL COMMENTS	The paper is really interesting and well written. It reports a deepen and systematic review of the use of Machine Learning in the early definition of fracture risk. This topic is really important, the osteoporosis is, in fact, considered the second most serious health issue. The work is not only a list of publications but is a critical analysis of all the pros and cons related to the reliability of the Machine Learning application in this field. The only lacking aspect in this paper is related to the innovative approaches that have been recently considered in some papers focusing on the micro-scale damage. It is, in fact, well known that bone structure and the consequent damaging are multi-scale. The micro-scale is considered the origin of the damage and the comprehension of the phenomena occurring at this scale is fundamental for an early diagnosis of the fracture risk. Anyway, to evidence micro-scale phenomena, advanced tools are needed. The most promising one, is the Synchrotron imaging combined with mechanical testing.

	However, synchrotron technique produces an enormous amount of data at an unprecedent resolution and there is the need and urge to define a validated tool to automatically analyze these sets of data. Artificial Intelligence is the tool that could overcame this issue. There are several papers focusing on these topics, among them the authors should cite: https://doi.org/10.1016/j.jmbbm.2021.104761 The fracture mechanics of human bone: influence of disease and treatment (ritchie), 10.1038/bonekey.2015.112 https://doi.org/10.1016/j.jmbbm.2022.105576 Fracture and Ageing in Bone: Toughness and Structural Characterization (ritchie): https://doi.org/10.1111/j.1475-1305.2006.00282.x https://doi.org/10.1016/j.engfracmech.2022.108582 https://doi.org/10.1016/j.mtla.2021.101229 https://doi.org/10.3390/ma14051240

REVIEWER	Myers, Thomas University of Rochester Medical Center
REVIEW RETURNED	11-Feb-2023

GENERAL COMMENTS	The authors did a nice job reviewing and presenting the current literature on this topic. My biggest critique is the conclusion of the abstract. I would suggest the authors temper their optimism for ML due due the lack of external validation. As they have correctly stated most (any?) studies lack external validation. Therefore, most of literature at this point is "proof of concept". We know ML works, we know that it works in medical literature. However, very few authors take it to the next level and externally validate their findings. The few studies in orthopaedics that have taken the external validation step see their AUC drop significantly. A few other points: 1. page 3 line 32 - I just want to be clear why not mentioning fracture risk was an excluding factor. It they have WHO definition of osteoporosis "<-2.5" isn't the fracture risk implied within the definition. I don't know, but this needs to be clarified.

REVIEWER	Kuo, Rachel Y L Botnar Research Centre, Nuffield Department of Orthopedics
REVIEW RETURNED	15-Feb-2023

GENERAL COMMENTS	Thank you for the opportunity to review your manuscript. This was an interesting systematic review and meta-analysis of the literature, on the prediction of fracture in patients with osteoporosis. I note that the authors have not included an analysis of adherence to current reporting standards. It would be interesting to see the results of such an analysis, as it would add complementary information to the quality assessment. For example, the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) could be appropriate, and is currently pending an AI extension. The authors have found a high proportion of papers at high risk of bias (64%). It could be an interesting exercise to perform a

	sensitivity analysis: comparing pooled results of all papers, to only those at low risk of bias, to only those at high risk of bias. This could give the reader an idea as to the influence (or not) of paper quality on reported model accuracy.

REVIEWER	Cook, Alex National University Singapore Singapore, Department of Statistics
REVIEW RETURNED	27-Apr-2023

GENERAL COMMENTS	I was asked to focus on the statistics in this paper. The results and the abstract are inconsistent in the presentation of sensitivity and specificity in training and testing data. The authors should check their numbers throughout in case there are other errors that have been missed. I don't really understand the logic of meta-analysing sensitivity and specificity, because most of the models will have created a range of both, and the authors of the original papers will have arbitrarily selected just one combination to report as their favourite. Using logistic regression as an example, the threshold to define someone as high or low risk can be moved up or down from 0 to 1, and the sensitivity and specificity will change accordingly. What then are the authors meta-analysing? The choice of threshold for low or high risk should be based on clinical need, which may differ from setting to setting and between use cases, after all. I am unconvinced in the finding that naïve Bayes performs best in training sets, because it is based on a single study. How are we to know that this study did not just lend itself well to distinguishing those at high risk? Figure 3 repeats the common failure to design forest plots for optimal understanding. The graphical element is squashed into a narrow strip, narrower than the space allocated to enumerating the models. Also, the labels for each point are visually separate from the graph so the eye cannot readily trace from the label SVM (say) to the point and CI. While I have focused as instructed on the statistics, I would recommend the whole paper be more thoroughly proofread before resubmission.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. John Carey, Galway University Hospitals, National University of Ireland Galway
Comments to the Author:
Thank you for asking me to review this article reviewing the predictive value of machine learning on fracture risk in osteoporosis.
The authors performed a substantial body of work, some of which is presented well. However parts are also confusing, and in my opinion this article would benefit from significant modification before being considered. It could be very interesting if it were shorter, clearer and focussed on the message in the title.
I have some major concerns about the authors understanding of the data and terminology used resulting in a lack of clarity and consistency for robust interpretation.

Response: We are extremely grateful that you take time out of your busy schedule to read our manuscript and give us these valuable suggestions. We appreciate your consideration about the presentation and clarity of the manuscript.
We acknowledge your concerns regarding the understanding of data and terminology in our article, which may have led to a lack of clarity and consistency. We have carefully addressed these concerns and modified the manuscript to enhance the interpretation of data.
We also appreciate your advice to shorten the article, make it clearer, and focuse on the message conveyed in the title. We have shortened the manuscript according to your suggestions so that the revised manuscript is more concise and effectively delivers the intended message.
Major Concerns:
Machine learning is a “method” or “technique”, not a “model” and this needs to be corrected and clarified throughout the paper. The model(s) generated will depend on the data, the characteristics of the study population and outcome of interest, the methods and analyses performed.
Response: We want to express the depth of our gratitude to you for your comments. We have revised related terminology in the paper. In the revised version, we have clearly differentiated between "machine learning methods" and "models," and ensure accurate descriptions of the relationship between machine learning methods and the models constructed based on variables such as research data, characteristics of the study population, and the outcome of interest. We have appropriately adjusted the wording in the paper to better reflect the relevance of machine learning methods and the generated models. However, some call it a model for the convenience of counting the number of a certain method such as ANN model used in previous scholars' research articles.
Applying the same ‘technique’ to different datasets will inevitably lead to the production of different results; if the results are robust however the differences should be smaller, and results more consistent. This needs due consideration here. In addition if the same ‘technique’ is applied to different datasets, it may result in different models being generated, particularly when there are different variables in the different datasets. This key aspect receives little attention and should be addressed as it will help explain some of the heterogeneity between studies.
Response: We appreciate your suggestion to consider the consistency of results across datasets. We completely agree with your observation that applying the same technique to different datasets can yield different results. It is important for the robustness of our findings. To address this concern, we have performed additional analyses to evaluate the robustness of our results. We acknowledge that different models may be generated when there are variations in variables across datasets. We have discussed this key aspect in our revised manuscript. We explain that the inclusion of different variables may contribute to the heterogeneity observed between studies. By addressing this issue, we provide a more comprehensive explanation for the variations in results. Once again, we sincerely appreciate your valuable comments, which are helpful for improving the quality of our manuscript. We believe that our revised manuscript now adequately addresses the concerns and provides a more robust and accurate interpretation of our findings.
The c-statistic has important limitations, particularly with respect to risk rather than discrimination, so this needs some discussion. This, and the over simplification using a Forest Plot in figure 3 needs further discussion. All these results are very dependent on multiple factors as outlined above, and also sample size, generalisability and frequency and accuracy of outcome/events.
Response: We completely agree with your point regarding the limitations of the c-statistic, specifically in terms of risk rather than discrimination. We have further discussed this aspect in the discussion section and highlighted the need for further research.
Additionally, we appreciate your comment about the potential oversimplification using a Forest Plot in Figure 3. We have revised this part to provide a more comprehensive explanation of the results to address this concern. Due to the magazine's limitation on the total number of charts, we put the previous figure 3 into the supplementary material.See Figures S1 and S2 for details.
Furthermore, we acknowledge the influence of multiple factors, such as sample size, generalizability, and the frequency and accuracy of outcome/events on our findings. We have emphasized these

factors in the discussion section, providing a thorough analysis of their potential impact on our findings.
The results as presented appear overly optimistic, as some studies have produced much lower values, e.g. Kong et al: PMID: 32161842.
Response: We are so thankful for your comment on the potential discrepancy in the reported results. The results of the study have changed due to the inclusion of 7 new papers, and we have revised this accordingly in the results and discussion sections.

Many clinicians are less familiar with artificial intelligence, machine learning and deep learning and a little more information on what machine learning actually is, how it differs from standard techniques, and the differences between supervised and unsupervised ML and why they matter is important for a journal such as this. Greater clarity is needed around this, for example a summary of how these various methods differ, perhaps a table? e.g. 'handles missing data' etc. Their work shows published results are markedly heterogeneous in rigour, design, technique and application and "Algorithmovigilance" is urgently needed (PMID: 33856479).
Response: In order to enable readers to understand the concept of machine learning, we have made corresponding modifications and supplements in background and discussion sections. We have added the validation methods of each model and the methods for dealing with the missing data in Table S4. We cite a review that details machine learning methods, which is basically divided into supervised and unsupervised learning. The review also describes the differences between them, which is not the focus of our research. Thus, we only briefly introduce machine learning methods in our article. The specific content is detailed in the following reference.
Gupta R, Srivastava D, Sahu M,et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 2021;25(3):1315-1360. https://doi.org/10.1007/s11030-021-10217-3
I am concerned the first sentence in the discussion section is misleading as this paper doesn't appear to include all currently available fracture risk prediction methods or models. In addition unless there are papers directly comparing different techniques and resulting model performance on the same datasets we cannot deduce which technique performs best. It would be better to consider studies which used a single method separately to those which compared different techniques in order to provide substance to this deduction such as the study by Kruse et al (reference 39) and then provide some sort of ranking, if of course that is possible. It may well be that different techniques perform very differently in different populations or on different datasets due to some of the reasons outlined in the discussion such as handling of missing data, smaller numbers, over or under representation of particular variables or outcomes and skewed/non-parametric predictors. The authors allude to some of these shortcomings in their discussion, but this could be more coherent and precise.
Response: Thank you for your valuable advice regarding the discussion section of our paper. We agree that direct comparisons between different techniques in the same datasets are necessary to determine the best performing method. Unfortunately, such comparisons were beyond the scope of this study. However, we recognize the importance of comparing techniques and will make efforts to perform relevant comparisons in future research.
Furthermore, we acknowledge the potential variations in the performance of different techniques across diverse populations and datasets due to the reasons mentioned in the discussion. We have described these shortcomings, emphasizing the potential impact of handling missing data, smaller sample sizes, and over- or under-representation of variables or outcomes. We also agree with your point that the discussion could be made more coherent and precise. In our revised manuscript, we have provided a clearer and comprehensive explanation of these limitations and their implications for future research.
In paragraph 7 of the discussion section the authors state 'existing ML models for fracture prediction focus on populations in Western countries' in fact of the studies they included 15 are European, 17 North American and 14 Asia-Pacific.

I don’t agree the data support their conclusion as their results suggest the field is very heterogeneous and sorely in need of a better structure and process. Surely the benefit of ML would be to accelerate the development of better models for specific populations rather than a ‘one size fits all’?
Response:According to your comments, we have revised the relevant content in paragraph 7, and the current research results have also changed due to the inclusion of 7 new literature references. Currently, the number of studies from the United States and Europe is equal, followed by studies from China. In general, the distribution is relatively even across the three continents. Therefore, the idea that there is some heterogeneity in the majority of studies from Western countries has been removed from the discussion.
We completely agree with your suggestion that ML can be used to develop better models for specific populations instead of relying on a "one size fits all" approach. In fact, we believe that the potential benefit of ML lies in its ability to accelerate the development of customized models for different populations, taking into account various demographic and geographical factors.
Minor Concerns:
There are many, so I only include some more salient ones here.
The authors could do with some independent help to review their grammar and terminology for this publication.
Figure 1. In box 3 there were 1673 records, n = 0 not retrieved, and in the next box (4) there are only 340 reports assessed! What happened to the other 1,433?? (No mention in the results narrative either).
The second paragraph of the “Results” needs to be rewritten / tidied up. The author state there were forty-six studies. Stick to numbers or words for consistency.
Then second sentence of this paragraph: “The majority of the studies were conducted in U.S. (n=10) and Canada (n=6),” needs to be reworded as this only represents 35% of the studies; ?most common location.
Next part of this sentence: “most were cohort studies (n = 40) or case-control studies (n = 6)” should read: “most were cohort studies (n = 40), and the rest were case-control studies (n = 6)”?
“Most study samples covered postmenopausal women (n=15)” is incorrect as this represents only 33% of the studies? Do you mean the number of subjects or the number of studies?
In my opinion it would be nice to include the sample size in table 1. This would be preferable to the author name as each study can be retrieved once referenced. This is included in Table S3 but I think Table 1 should include this, and look neater and tidier like tables S3 and S4.
Consider deleting the word “data” from the rows in the column titled “Data Source” and put in things like “Database”, “Registry”, “Medical Records” etc.
The wording used in the next column is very confusing under “Sample Population Type”.
I am unclear what the difference between a “patient”, “older women”, “postmenopausal women”, “older men”, “inhabitant”, “elderly” etc. is. Usual terms include: “men”, “women”, “patients”, “subjects”. Please simplify and clarify.
Consider substituting “mean” for “average” for clarity in the Age column title.
In the “Fracture site” Column would it not be better to just list as “hip”, “vertebral”, “multiple” rather than repeatedly including the word “fracture”?
The “ML Models” are not really “models” but “methods”. The final model will be derived using these various machine learning methods, e.g. LR, ANN etc. Very little data is presented on the ‘models’.
Paragraph 4: “We roughly classify”. Please delete “roughly” and clarify. I am unclear why fracture history and falls are included with “demographics”, while “osteoporosis” and “fracture type” are included under “Comorbidity”. This needs tidying up. Recommend keeping demographics to demographics, and clinical risk factors and comorbidities together. Please clarify if bone mineral density is in ‘grams/centimeter2’ or the ‘T-score’, or both.
Figure 2 appears incorrect:
1. “Low” is not included in the “overall” bar (5th column/bar)
2. There is a yellow square box in the figure legend with no wording beside it, Omit??

While I like the concept of a Forest Plot to summarise lots of data, I am concerned figure 3 is difficult to understand/interpret. The authors evaluating ML ‘methods’ to develop ‘models’ but provide no information on what was actually in these models, or how they compare to any other model, e.g. FRAX. A plot like this may have some value if it only included results from ‘validation’ datasets, though some of the same concerns remain.
Response: We are much obliged to you for your very detailed and valuable comments on our paper, and we are very sorry for these errors in our manuscript. We have revised these issues one by one in the revised version, especially the content in Table 1. Due to the excessive content of the tables, Table 1 in the previous version has become Supplementary material Table S3. We also added validation methods and model evaluation metrics in Table S3. We redrew the previous forest map and now put it in supplementary material Figures S1 and S2.
In addition, we have streamlined paragraph 4 and annotateed BMD (g/cm2) in the endnotes in Table 1.

Reviewer: 2
Dr. Laura Maria Vergani, Politecnico di Milano
Comments to the Author:
The paper is really interesting and well written. It reports a deepen and systematic review of the use of Machine Learning in the early definition of fracture risk. This topic is really important, the osteoporosis is, in fact, considered the second most serious health issue. The work is not only a list of publications but is a critical analysis of all the pros and cons related to the reliability of the Machine Learning application in this field.
The only lacking aspect in this paper is related to the innovative approaches that have been recently considered in some papers focusing on the micro-scale damage. It is, in fact, well known that bone structure and the consequent damaging are multi-scale. The micro-scale is considered the origin of the damage and the comprehension of the phenomena occurring at this scale is fundamental for an early diagnosis of the fracture risk.
Anyway, to evidence micro-scale phenomena, advanced tools are needed. The most promising one, is the Synchrotron imaging combined with mechanical testing.
However, synchrotron technique produces an enormous amount of data at an unprecedent resolution and there is the need and urge to define a validated tool to automatically analyze these sets of data. Artificial Intelligence is the tool that could overcame this issue.
There are several papers focusing on these topics, among them the authors should cite:
https://doi.org/10.1016/j.jmbbm.2021.104761
The fracture mechanics of human bone: influence of disease and treatment (ritchie),
10.1038/bonekey.2015.112
https://doi.org/10.1016/j.jmbbm.2022.105576
Fracture and Ageing in Bone: Toughness and Structural Characterization (ritchie):
https://doi.org/10.1111/j.1475-1305.2006.00282.x
https://doi.org/10.1016/j.engfracmech.2022.108582
https://doi.org/10.1016/j.mtla.2021.101229
https://doi.org/10.3390/ma14051240
Response: We would like to express our heartfelt appreciation for your comment on the paper. We also appreciate that you point out the importance of considering micro-scale damage in the early definition of fracture risk. We agree that the understanding of micro-scale phenomena is crucial for an accurate diagnosis.We apologize that we did not cite references on the use of synchrotron imaging and AI in micro-scale damage analysis. After downloading and reading these documents, due to the limitations of the word count and the number of cited references, we finally added reference 1 and 2 (A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications;Deep learning approach to assess damage mechanics of bone tissue) to our revised version.Thank you for bringing this to our attention.

Reviewer: 3
Dr. Thomas Myers, University of Rochester Medical Center
Comments to the Author:
The authors did a nice job reviewing and presenting the current literature on this topic. My biggest critique is the conclusion of the abstract. I would suggest the authors temper their optimism for ML due due the lack of external validation. As they have correctly stated most (any?) studies lack external validation. Therefore, most of literature at this point is "proof of concept". We know ML works, we know that it works in medical literature. However, very few authors take it to the next level and externally validate their findings. The few studies in orthopaedics that have taken the external validation step see their AUC drop significantly.

A few other points:

1. page 3 line 32 - I just want to be clear why not mentioning fracture risk was an excluding factor. It they have WHO definition of osteoporosis "<-2.5" isn't the fracture risk implied within the definition. I don't know, but this needs to be clarified.
Response: Many thanks for your comment on the conclusion of the abstract. We appreciate your point about the lack of external validation in the current literature on machine learning (ML) for fracture risk assessment. You raise a valid concern about the need for external validation to ensure the reliability and generalizability of ML models in this field.We have revised the conclusion of the abstract to reflect the need for further research that includes external validation to assess the true predictive power of ML models. “ However, most current studies lack external validation. Therefore, future research is needed to validate and improve the existing predictive models for osteoporosis risk rather than developing new models.”
We want to express the depth of our gratitude to you for your question and suggestion regarding page 3, line 32. We appreciate your attention to details and your suggestion on clarifying why fracture risk was an excluding factor in our study. You are correct that the WHO definition of osteoporosis includes a bone mineral density (BMD) value of "<-2.5." This value signifies a significant decrease in bone density and implies an increased fracture risk. However, in our study, we chose to explicitly mention fracture risk as an excluding factor to ensure that only studies on fracture risk prediction using machine learning were included in our review. All search terms related to "fracture" are specifically shown in Table S2, and we obtained relevant studies based on literature search.

Reviewer: 4
Dr. Rachel Y L Kuo, Botnar Research Centre
Comments to the Author:
Thank you for the opportunity to review your manuscript. This was an interesting systematic review and meta-analysis of the literature, on the prediction of fracture in patients with osteoporosis.

I note that the authors have not included an analysis of adherence to current reporting standards. It would be interesting to see the results of such an analysis, as it would add complementary information to the quality assessment. For example, the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) could be appropriate, and is currently pending an AI extension.

The authors have found a high proportion of papers at high risk of bias (64%). It could be an interesting exercise to perform a sensitivity analysis: comparing pooled results of all papers, to only those at low risk of bias, to only those at high risk of bias. This could give the reader an idea as to the influence (or not) of paper quality on reported model accuracy.

Response: We are very/extremely grateful for your valuable comments and suggestions to include the analysis of adherence to current reporting standards and perform a sensitivity analysis based on the risk of bias. We agree that analyzing adherence to current reporting standards, such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement, would provide important complementary information to the quality assessment. We have assessed the adherence to reporting standards by using CONSORT (consolidated standards of reporting trials) for randomised studies and TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) for non-randomised studies. Risk of bias was assessed by using the Cochrane risk of bias tool for randomised studies and PROBAST (prediction model risk of bias assessment tool) for non-randomised studies. (PMID: 32213531)
The PROBAST to assess the risk of bias in the included original studies is more suitable for our meta-analysis of risk prediction models. This tool includes a large number of questions in four different areas: participants, predictive variables, outcomes, and statistical analysis, reflecting the overall risk of bias and overall usability.
We appreciate your valuable suggestion to perform a sensitivity analysis in our study, comparing the pooled results of all papers to those at low risk of bias and high risk of bias. This analysis would provide valuable insights into the influence of study quality on the reported model accuracy, as you mentioned. We have performed the sensitivity analysis as recommended (Supplementary material Figure S3-S8).

Reviewer: 5
Alex Cook, National University Singapore Singapore
Comments to the Author:
I was asked to focus on the statistics in this paper.

The results and the abstract are inconsistent in the presentation of sensitivity and specificity in training and testing data. The authors should check their numbers throughout in case there are other errors that have been missed.

I don’t really understand the logic of meta-analysing sensitivity and specificity, because most of the models will have created a range of both, and the authors of the original papers will have arbitrarily selected just one combination to report as their favourite. Using logistic regression as an example, the threshold to define someone as high or low risk can be moved up or down from 0 to 1, and the sensitivity and specificity will change accordingly. What then are the authors meta-analysing? The choice of threshold for low or high risk should be based on clinical need, which may differ from setting to setting and between use cases, after all.

I am unconvinced in the finding that naïve Bayes performs best in training sets, because it is based on a single study. How are we to know that this study did not just lend itself well to distinguishing those at high risk?

Figure 3 repeats the common failure to design forest plots for optimal understanding. The graphical element is squashed into a narrow strip, narrower than the space allocated to enumerating the models. Also, the labels for each point are visually separate from the graph so the eye cannot readily trace from the label SVM (say) to the point and CI.

While I have focused as instructed on the statistics, I would recommend the whole paper be more thoroughly proofread before resubmission.
Response: We want to express the depth of our gratitude to you for your insightful comments regarding our study. We acknowledge the inconsistency in the presentation of sensitivity and specificity in the results and abstract. We apologize for any confusion caused and appreciate your suggestion to thoroughly check the numbers throughout the manuscript to ensure accuracy an

consistency. We have carefully reviewed the data and made necessary corrections to ensure consistency in the presentation of the sensitivity and specificity values.

We do appreciate your concern about the meta-analysis of sensitivity and specificity. It is true that the selection of a specific threshold to define high or low risk may cause variations across different studies. We agree that the choice of threshold should be based on clinical need and may differ between settings and use cases. In our study, we performed a meta-analysis to estimate the overall performance of various risk prediction models, given the reported sensitivity and specificity values across different thresholds. While this approach may not capture the full range of sensitivity and specificity values for each model, it provides a comprehensive overview of the performance across multiple studies.

We acknowledge the limitation of having only one study supporting that naïve Bayes performs best in training set. We added seven newly retrieved papers to our study, and the results changed accordingly. The deep learning algorithm performs relatively well now.

We appreciate your comment on Figure 3. To enhance clarity and improve readability, we have redesigned the forest plot to allocate more space for the graphical element and ensure better visualization of the model points and corresponding confidence intervals. We redrew the previous forest map and now put it in supplementary material Figures S1 and S2.

Lastly, we appreciate your recommendation to thoroughly proofread the entire paper before resubmission. We have meticulously proofreaded the manuscript to address any remaining errors and ensure the overall quality of the manuscript.