Peer Review Report

Review Report on Genetic Risk for Osteoporosis and the Benefit of Adherence to Healthy Lifestyles

Original Article, Int J Public Health

EVALUATION

Q 1 — Please summarize the main findings of the study.

In this study, the authors construct a genetic score for bone mineral density, a composite "healthy lifestyle" score based on environmental variables, and show in detail how these variables are associated with risk of fractures or osteoporosis.

Q 2 — Please highlight the limitations and strengths.

The chosen methods are standard in the field, the dataset is appropriate, and the analysis was conducted well from the technical side, with various relevant outputs and diagnostics presented. The reported associations can be valuable to the field. However the main goal of this study is not clear: a lot of claims about causality and better prediction are made, and either of those goals would require deeper analysis than presented here. Various minor writing issues are present.

Q 3 — Please provide your detailed review report to the authors. The editors prefer to receive your review structured in major and minor comments. Please consider in your review the methods (statistical methods valid and correctly applied (e.g. sample size, choice of test), is the study replicable based on the method description?), results, data interpretation and references. If there are any objective errors, or if the conclusions are not supported, you should detail your concerns.

My major question is the goal of this study. In the Introduction, the goal is set to be better prediction of the outcomes, but this is completely ignored further: no direct prediction metrics are shown in the analysis, such as PPV, F1 score, R2, average time to event or similar, no comparison with accuracy of existing methods is provided, and the chosen methods are not optimized for prediction (very simple regression models, or Cox model which does not estimate the baseline and so is not directly useful for prediction).

In contrast, the Discussion and Conclusions are mostly about causal inference: the associations of the selected lifestyle factors are discussed in causal language, some causal mechanisms are proposed, and public health recommendations made. The current study design is not all appropriate for making new causal claims. The models include very few covariates with unclear selection criteria, no comments on the validity of the measures (in particular the dietary ones) and other epidemiological aspects that are necessary to separate causality from confounding.

Finally, based on the frequent mention of "joint" or "combined" effects in the text, I was expecting that the study might focus on testing the interaction between some known causal factors and the genetic risk scores, but no such test was done. The authors might consider if this may be of interest, as it seems fully feasible within this data.

In any case, additional analyses probably will be needed as appropriate for the chosen goal. Particular care is needed to remove any new causal claims if the analysis is not adapted to support those.

There are some minor issues with descriptions and many smaller typos or writing issues which I list below.

Abstract. Very abbreviations-heavy, consider expanding some or rephrasing. In particular OP, eBMD, HR are not defined.
Throughout the abstract, it should be more explicitly stated what trait the GRS is constructed for – at least state it explicitly in line 13 when it is introduced.

l. 32. "significant loci associated" – maybe rather "loci significantly associated", or just "loci associated".

l. 41-50. This entire paragraph is confusing. GRS seems to be compared with some other prediction tool, but not clear with what: "GRS can identify a greater proportion of people with disease risk", "GRS constructed by multiple genetic loci could improve the predictive accuracy".
"These studies using GRS have been effective in predicting OP risk" – seems to contradict the above. The cited study in ref. 18 predicted childhood BMD change, not sure how relevant that is to the OP risk.

l. 54. "Their combining effects" – maybe "the combined effects"?

l. 88-89. Could you clarify what ancestry was used (British or European) and how it was determined (are these self-reported values from the questionairre?).

l. 91-92: A short explanation about the meaning of the SOS and BUA variables should probably be added – at least to help readers less familiar with this phenotype.

Fig. 1: "Individuals with Successful Genotyped" – "Individuals successfully genotyped"
"Not current smoke" – "not current smoker" or similar
"Modern drink" – "Non-excessive drinker"?

l. 106. "Similar analyses were conducted in different gender groups" – I understand this to mean that the analyses were stratified by gender, but next sentence says that gender was included as a fixed covariate. Please clarify.

l. 117. Alcohol lifestyle variable defined very differently here (drinking 3 or more times per week) and in table S2 (drinking once per month). Please fix.

l. 120. "Having three of the seven food groups" is not clear, please define in more detail what are "the seven groups" and what amounts to "having" one. Furthermore Table S2 says "five of the seven groups" for the same criterion. Table S2 is clear and helpful, but the main text should also clarify that the "healthy diet" variable requires both the specified servings, and drinking milk.

l. 126. This index does not seem to have been used anywhere, only the individual factors or a 3-level grouping of them.

l. 129-132. I think the description of the procedure here would not be clear for a reader unfamiliar with the PLINK procedures specifically. I would not say that PLINK's --clump "singles out independent SNPs"; rather, it combines SNPs into associated regions, with the lead SNP for each simply chosen by the highest significance. The sentences that follow refer to "the corresponding SNPs", and "the lead SNP generated from GWAS analysis", and I think it could be made more clear that these are the lead SNPs for each clump.

l. 138. The models used here and for creating the GRS are slightly different: mixed models were used before, and different covariates added. Is there any particular reason for these differences? Additionally, suggest stating the software and packages used for the analyses in this paragraph.

l. 140. "Using multivariate ... to demonstrate the risk of OP." Sentence seems wrong or incomplete, please fix. Also "multivariate" means many outcomes – presumably multiple or multivariable?
l. 142. Incidence of cases, not of people.
l. 143. The residuals cannot be significant or not. The proportional hazards test typically tests whether there is an association between these residuals and time. The procedure admittedly is not simple to describe – I

recommend stating the program, function and parameters used to conduct the test, which should be sufficient for understanding and reproducing the analysis.

l. 144. COX should be Cox. The category "unavoidable" should be "unfavorable" based on the text further. How did the lifestyle enter the analysis: as another covariate (continuous or factor?), or was an entirely separate model fit in each of the strata? Was there an interaction between them? Fig. 3 caption says "stratified", which typically implies a different model with a free baseline hazard for each level, but then the figure shows hazard ratios for each level, which contradicts that.
It is not clear what, if any, other covariates were used in the Cox model.

l. 151. Why is the BUA variable specifically mentioned here – I assumed so far that quality control used many variables, as indicated in Methods and Fig. 1?

l. 155. The numbers do not match in various ways. The percentages here imply a total size of the test set of about 75000. Fig. 1 says that the test N=84000. Table 1 is labelled as "baseline characteristics in the UK biobank cohort" and also as "test set", but lists characteristics of N=100545, which seems to be the test +selection set. Most importantly, the text says that fracture frequency was about 5 %, whereas the table shows it was close to 15 %. I am not sure if something happened with the analysis, or just the labelling, but all this needs to be thoroughly fixed.

The distribution of BMD is also important to see, as it is the main outcome of the GWAS part of this study. Could be included as additional summary statistics in Table 1, or ideally as a separate figure showing the full distribution of it in the different case and control groups.

l. 171. Three levels were defined "according to quartiles" – this should be described fully here, since if the levels were just the quartiles, I would expect four. Table S4 caption needs to be fixed, as it says that the levels were based on quintiles.

Fig 2 caption: a lot of typos and inappropriate capitalization. Why is the "prospective analysis" term suddenly introduced for subfigures E and F?

l. 178. "An approximate gradient of reduced OP risk... was found" – I am not sure what this means. Maybe something like "the risk was smoothly decreasing with increasing GRS" would be more fitting?

l. 181. The trend analysis, spline model fitting and the significance test for non-linearity all should be described in the methods, as the choice of tools for these tasks is not obvious. The methods also mention a test for proportionality of hazards which does not seem to be used here?..
l. 182: "P_no-liner" – fix spelling.

l. 188. Claiming that these lifestyles are "beneficial to the bone" is too strong, and not supported by the brief literature summary conducted in the methods. The two reviews cited for diet show mostly no evidence for its effects on OP, and also not for the particular diet definition from this study. The study cited for alcohol intake in fact shows protective effects of drinking, contrary to its treatment here. The study cited for sunshine exposure did not analyse OP or bone fractures at all.
To be clear, I agree that these are reasonable choices of factors to investigate in this study, but the authors should not present them as established causal factors.

l. 202. "We defined three levels of healthy lifestyle groups based on the eBMD results with the highest statistical power." I do not understand what power is referred to here, i.e. what is being tested and how? Generally, I would guess that in most cases keeping more groups for a variable preserves higher power, so I do not understand why this re-grouping was at all needed.

l. 215. The "dose-response manner" here seems superfluous – the results simply show the effects of the covariates on the risk of OP or fracture, unless I am missing something.

Table S7 caption again incorrectly states that quintiles instead of quartiles were used for defining GRS groups. Also the table appears to have exactly the same information as figure 3, with the addition of the trend p-values, so maybe those could simply be added to the figure?

l. 222–230. Example of the goals and discussion not matching: lots of causal language ("risk can be reduced by advising the public..."); focus on the "combination" of GRS and environemental factors, even though both seem to have additive effects and their interaction was not tested.

l. 261–266. Again some unclear and inappropriate comments on causality. Are the authors implying that cohort design is needed for establishing causality? What are the findings that are made "more confident" by causality? What is "basic data for OP prevention"?

l. 275. Not sure what bias is referred to here – some more explanation is needed, and in particular how it is offset by sample size.

l. 277. The last sentence seems incomplete.

PLEASE COMMENT

Q 4 **Is the title appropriate, concise, attractive?**

The title is concise, but implies causality of the identified associations ("benefit of adherence to...") and should be changed unless the authors significantly expand the study for valid causal inference.

Q 5 **Are the keywords appropriate?**

Yes

Q 6 **Is the English language of sufficient quality?**

A lot of small language issues are present, but they generally can be easily fixed and do not interfere with reading

Q 7 **Is the quality of the figures and tables satisfactory?**

Yes.

Q 8 **Does the reference list cover the relevant literature adequately and in an unbiased manner?)**

Yes

QUALITY ASSESSMENT

		Rating
Q 9	Originality	2/5
Q 10	Rigor	2/5
Q 11	Significance to the field	3/5
Q 12	Interest to a general audience	4/5
Q 13	Quality of the writing	2/5

Q 14	Overall scientific quality of the study

REVISION LEVEL

Q 15	Please make a recommendation based on your comments:

Major revisions.