nature portfolio

Peer Review File

Metabolomic and genomic prediction of common diseases in 700,217 participants in three national biobanks

Corresponding Author: Dr Jeffrey Barrett

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I thank the authors for their response to my review regarding characteristics of excluded subjects from the biobanks, additional details re: LASSO models, and adjustment of statements related to proteomics. I am struggling a bit with the other aspects of the authors' response as it is collectively wrapped up in what is described as a 'philosophical difference' in objectives which are now no longer focused on demonstrating improvement in risk prediction over conventional approaches but rather some sort of 'proof of principle' that routine NMR testing can identify individuals at high risk of multiple disease and enhance a "national preventive health check program". The difficulty is the original intent seemed to have been indeed to make this claim (as per the original abstract and manuscript submission) and the revised manuscript continues to promote this idea even if the claims have been scaled back. Importantly, this message is being delivered by an investigative group that is dominated by employees of the company that sells the NMR test. Given these circumstances, I think it is critical to ensure that statements and conclusions made in this manuscript are unequivocally supported by the data and analyses. For this reason, I feel compelled to be more prescriptive in my critique this time around.

First, I recommend the authors further scale back and completely refrain from any commentary or speculation on whether the results of the study provide adequate evidence that NMR omics testing can be incorporated into public health practice to enhance national preventative check programs through screening testing performed on a regular basis. The analytic approach and the results arguably do not rise to that level of evidence for a several reasons (also echoed by reviewer 2): 1) the NMR scores have not been tested against the most advanced clinical scores/predictors and, for some of the diseases presented, in the correct clinical setting (see below) 2) the analyses do not correct for effects on scores that could be a consequence of the use of medications (while the authors present some compelling evidence for why statins do not artificially enhance discrimination for MI, they do not present results adjusted for statin use and do not systematically address a multitude of other drugs that could be preferentially used among cases or controls) 3) the statistical comparisons have been anchored on comparisons of hazard ratios rather than AUCs or other metrics of discrimination.

Along these lines, the authors themselves refer to a UK Biobank preprint in their discussion that tests the value of proteomics profiling in the UK Biobank (Gadd et al., reference 31 which now has a version 2 preprint). That manuscript adopts an approach to testing -omics versus clinical prediction that is in line with comments above and ideally should be emulated here. That manuscript also steers away from any discussion about implementing proteomics in clinical practice. Until a more comprehensive approach is adopted, I think the final sentence of the abstract should be removed or scaled back considerably.

In this context, I encourage the authors benchmark the metabolomics platform against more 'complete' or established clinical algorithms tailored to the disease whenever possible if emulating the approach adopted by the proteomics group is too much of a rewrite. These additional analyses should not be burdensome as they simply entail a rerun of scripts adding or substituting additional covariates to the base / clinical models in the same manner pack year smoking was considered. New data could then be added to the current figures and AUCs tables. For example, for diseases such as MI and ischemic stroke, the 'clinical' benchmark could be the QRISK3 score itself (restricted to the age range for the score) in the UK Biobank, or the inclusion of as many variables as available from the score in the baseline model. For diabetes, it could be the QDiabetes score. For COPD, lung cancer, and alcoholic liver disease, and cirrhosis, it would be helpful to add analyses in the figures and tables restricted to ever smokers and ever drinkers (some of this is in the text already).

The discussion section is not optimally written. The first paragraph section about AUCs and clinical risk scores is confusing. The first few sentences of the final paragraph re: incorporating NMR metabolomics in public health practice are premature and should be avoided for reasons described above. Lastly, a comprehensive paragraph describing the limitations to these analyses is missing (again, see Gadd et al. for a good example to emulate). For several diseases presented, it simply is not possible to test against the current 'gold standard' approach to preventing or identifying specific disease early. For example, comparing NMR testing vs. screening with a CT scan for lung cancers among smokers is likely not possible to do with the data at hand. The same goes for 1) colon cancer and fecal occult blood tests and/or colonoscopies, 2) more sophisticated screening questionnaires for unhealthy alcohol use (e.g. AUDIT/AUDIT-C, CAGE), 3) COPD and pulmonary function testing among smokers. The discussion section needs to be more explicit about these limitations of the analytic approach presented (and the data available). Re: alcohol liver disease and cirrhosis, is it possible that the NMR testing is generating a signature of active heavy drinking? If yes, this should be explicitly mentioned in the manuscript as this may have major implications when trying to implement this test, some of these ethical.

Table 2 is a nice addition of AUC but can be improved. I am not clear if the first three columns have any clinical variables. I think they are PGS and metabolomics or combined alone but it would be helpful to make that clear in the table as a footnote. Clinical should probably be presented as first column. Some diseases should have more than one entry presenting comparisons to more complex models as suggested above. Also, why not include clinical+PGS, clinical + omics/PGS?

Reviewer #2

(Remarks to the Author)
Thanks for the opportunity to review this manuscript. The authors test the predictive utility of 36 metabolic biomarkers assayed using the Nightingale platform in ~477,000 individuals across three cohorts. The authors further compare risk stratification by metabolomic scores vs. polygenic risk and, as expected, observe greater hazards associated with high metabolomic vs. high polygenic risk. Incorporating metabolomics appears to improve prediction to varying extents across disease outcomes. Overall, the analysis has several strengths and suggests potential clinical utility of these biomarkers in practice. Although I agree with the authors that each disease needs its own careful consideration in future analyses, and although the authors appear to be de-emphasizing improvement in risk prediction in this revised submission, I think the manuscript would benefit from a bit more demonstration of incremental utility and new insights compared with previously published work examining Nightingale metabolomics in the UK Biobank and elsewhere, especially given the 'biomarkers-for-all' framing of Discussion.

As noted above, although I agree with the authors that each disease needs its own careful consideration and that doing this for all diseases may be beyond the scope of this paper, it would be helpful and informative to 'dig into' one or a couple of use cases where robust risk prediction approaches are developed and already in clinical practice. For example, as noted by the authors, the QRISK and FINRISK scores are validated for ASCVD outcomes (coronary heart disease events and ischemic stroke). I suggest the authors provide delta AUC and net reclassification indices for incident myocardial infarction and ischemic stroke when applying the metabolomic score on top of QRISK/FINRISK to statin-untreated patients (i.e., the populations where these scores are used to inform statin prescribing decisions); this would reasonably simultaneously address several of the comments raised by previous Reviewer 2 in response to the prior submission.

Table 2: It would be helpful for the authors to provide AUCs for clinical score + PGS and, separately, for the clinical + metabolomic score (in addition to clinic + combined omics), as well as P-values for comparisons of each approach to the clinical-only score (e.g., using the Delong-Delong test).

Minor:

Unless I am missing it somewhere, the actual metabolites used in the score for each disease outcome are provided only in Table S3, which makes it difficult to ascertain how the scores are similar/different across conditions. Can the authors provide this info in a separate, easier-to-read supplementary table?

The authors should add P-values for deltaAUC to Table S5.

It appears there may be an error in Figure S4b – the forest plot appears duplicated, whereas the headings "Myocardial infarction" and "Ischemic heart disease" each apply only to a subset of the ICD codes currently listed under both headings.

The authors might consider highlighting for the reader who is more peripherally familiar with the multi-omic biomarker literature (e.g., in the Discussion) how the present analysis is distinct from/additive to the Buergel et al., Nat Med 2022 paper.

Reviewer #3

(Remarks to the Author)
In this paper the authors trained risk prediction scores from NMR based metabolomics features and polygenic risk scores for 12 diseases selected based on these being leading causes of disability-adjusted life years in high-income countries. In half of the UKB data Cox proportional hazards models were applied to predict incidence of the 12 diseases based on feature selection by Lasso (and five-fold cross-validation) from 36 clinically validated blood based metabolomics features.
Using the weights estimated in UKB, the other half of UKB and the Finish and Estonian biobanks were used to test the relative and joint performances of the metabolomic scores and PGSses for predicting the 4 years incidence of 12 diseases,

and 10 years prediction across follow up compared to clinical markers as well as the 10 years performance of repeated metabolomic measures at different time points.
This is an very interesting study of large scale metabolomics and genetics data in three large national biobanks. Many studies aim to generate risk predictors from omics data and the strength of the current study is the large scale analysis allowing a comparison to genetic risk factors (PRS) and clinical risk factors in three studies. The high-risk groups showed consistently increased risks in all biobanks (3 countries, across plasma vs. serum sample types, various enrolment criteria, and fasting protocols).
The ultimate goal of the study is not clearly formulated. It does not seem to be to test for the optimal predictive information for these DALYS in the 249 metabolomics features as compared to specific clinical risk markers for these diseases, which is a shame given the large dataset. The focus seems more on whether 36 clinically validated features (out of 249) can ultimately be used in real-world public health settings. Authors should clarify the goals, because in a dataset as large as the one described here one would expect a search for the optimal prediction from available features as well, or if focused on 'real-world public health settings' on meeting or improving the gold standard. But what is the standard for clinical risk estimation in the clinic for these DALYs, the only hint is the last paragraph indicating a comparison to NHS health check. Is that the ultimate aim ?
The clinical score is generic, based on age, sex, smoking pack-years, BMI, systolic blood pressure, HDL cholesterol, total cholesterol, and weekly alcoholic drinks. Are these the variables used in the clinic for all DALYs ? Justification and reasoning on the clinical scores and variables used for comparison of prediction of these 12 DALYs must be better presented. One can imagine that missing from the clinical variables is CRP or albumin, also regularly used in the clinic, and also in commercial health checks; and certainly present in the cohort studies, but why then not included here for comparison.
Another point is : there is comparison to clinical prediction across 10 years trajectories but not across the 4 years that the paper starts with, would that not be also relevant for public health ?
The result section does not summarize much on differences between the cohorts (except collection) and yet in Figure 4 they show different results here and there so do discuss Table 1/S2 a little better. (differences in health in the three studies, which may be relevant for discussion Fig 4).
As indicated it can be more clear from the start of results that all scores were based on 36 clinically validated markers out of 249 available ones measured in total.
It must be shown in supplements which features are most prominent in the various scores. One can not see which features are most informative, which seems especially interesting for the four conditions for which metabolomics adds to clinical risk prediction. And also to understand why scores for COPD, Lung cancer and depressive disorders cluster as can be seen in Figure S6.
The goal of future rapid translation seems a bit in the way of trying to get the best predictions which should ideally have been part of this epidemiological study, is the maximum for this platform reached when using the 36 features because the rest is redundant or have studies shown that the 36 only get halfway as compared to the unused features ? It is a logical analysis to explore to what extend other features could have improved prediction. In other metabolomics work usually about 65 independent features are used for feature selection, the choice for the 36 is based on allowing standardized future application of the score, not to test the most optimal prediction. But which metabolites are missing if you compare the 36 to the metabolites strongly associating to outcomes in the papers referred to about previous work (15-18), do you expect to gain predictive power if these were added in future.
Table 2 is highly relevant when considering the translational value of the scores and platform.
Some observations can be phrased more clearly I think
1. Considering the 10 years risk prediction, omics scores based on the 36 features (27 separate features and 9 ratios) used here, only exceed the clinical prediction above marginal for T2D, MI, and the liver conditions. And that clinical prediction does not even include albumin or CRP. Leaves the reader with the feeling: could prediction have been increased when more measured features were used. Table 2 shows that for 7 out of 12 diseases, the prediction improvement as compared to the clinical variables is marginal (.01 increased AUC). For Colon cancer this is slightly higher .02) and for MI, T2D, and the liver prediction is really improved.
2. Clear is that genetic scores hardly contribute to the improvement that was observed. In the four improved cases it is mainly (MI) or only (T2D and liver diseases) metabolomics that increases prediction..
3. From a translational point of view it is highly valuable that the variable weights estimated in the UKBB could be used in other studies without normalizing the biomarkers within each study separately.
4. So one may conclude that metabolomics alone based on the clinically validate features generates the same information as the clinical score for most DALYs and improved it in four conditions
5. But then again it is critical what is currently used and would clinic and public health agree on the clinical variables selected here.
It is clear that the metabolomics scores from the 36 features selected are not specific for these DALYs, but neither would the clinical score be considering the selected variables. Please include the extent to which the clinical features relate to the DALYS Since the predictors are not very specific, how can they be translated, were the clinical predictors less correlated ?
The analysis of repeated measures is very interesting indeed; could authors discuss whether there may be impact of excluding disease events between baseline and repeat visits but leaving the persons in. Persons could be missing in the repeat measure because they were unhealthy at baseline, had their event and did not pay the repeat visit or died; these would be left out. So those surviving the event and paying the next visit are in the second visit data and then may have no event anymore unless a recurrent one. So, for those who had an event maybe treatment lowered their metabolomic score.
The abstract mentions only that lifestyle could have changed the score but what was investigated only explained a small portion of the change. The sentence ' People whose scores change have different future risk of disease, suggesting that repeat measurements capture the benefits of lifestyle change' would have to be rephrased in something like: repeat measures capture changes bot health statis and future disease risk possibly due to treatment, lifestyle changes or other beneficial factors.
The discussion touches proteomics and age of the study populations very briefly and no mentioning of how the

metabolomics scores may be improved by the other available metabolites ot other metabolomics technologies also used in epidemiological studies (Mass spec based lipidomics, Metabolon platform). The proteome was studied in UKBB , so do authors foresee specific proteins that may add to predictions ? What are expectations.
Because participants are still young (also for relating to frailty by the way) , perhaps the predictions improve in older persons where generic scores representing health status can be useful not only for lifestyle improvement.
Overall the paper is very interesting by its large scale, diversity of cohorts, comparison of two omics PRS and showing that the validated features included in the score are robustly indicating increased risks in all biobanks (3 countries, across plasma vs. serum sample types, various enrolment criteria, and fasting protocols) holding promise for using metabolomics as generic health check to support preventive health care in the population at large. It can be more clear in showing whether and how predictions may be improved, and perhaps a more realistic comparison with clinical variables as they are measured in real life situations for these DALYs in clinic or population. Your study has to be convincing that once blood sampling would be included in population wide health checks, the metabolomics will exceed prediction over clinical variables that would then include common blood based markers.
It should also be pressed that it is of great interest to the biomarker field to explore if one can turn affordable, robust omics biomarker assays into practically usable tests and the paper is a serious effort in testing this out.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I would like to thank the authors for their modifications and for including additional data. It's quite remarkable that the entire UK biobank has now been profiles with the metabolomic platform. I think overall the manuscript has been strengthened substantially. However, I continue to feel the need to be prescriptive to manage potential conflict of interest. Thus, I have these very specific suggestions to make one last time. A vast majority of these are simple changes to the text but a couple may require new simple analyses to clarify.

ABSTRACT: The abstract highlights positive results for all comparisons that are minimally adjusted but does not summarize the results of the critical last set of analyses comparing clinical scores to clinical + metabolomic scores. I may suggest replacing the current final sentence of the abstract with the following adjustment: "Lastly, we assessed the incremental predictive value of metabolomic scores over existing clinical risk scores for multiple diseases and found modest improvements in discrimination for several diseases whose clinical utility, while promising, remains to be determined." I believe this summary of findings is extensively described in the discussion section already but is noticeably absent in the abstract.

TEXT: The term "future risk" and "future disease onset" is used throughout the manuscript. I do not think the word "future" is needed. It is redundant.

I am somewhat dissatisfied with multiple section headings and legend titles to the figures. They are either not descriptive enough or suggesting something more than what is being delivered in that section. In multiple section headings, the term "performance" or "risk prediction performance" is used, when performance is not being formally assessed at least not against the 'gold standard'. Below I make suggestions for changes in titles of sections and figures. Within some sections I also make additional suggestions for change.

Please change section title "Risk prediction performance" (Page 3, line 25) to "Baseline age and sex minimally adjusted metabolomic scores and incident disease"

I would change the next title from "Investigations into the scores" to "Sensitivity and Subgroup Analyses of minimally adjusted scores"

There is no formal comparison of performance being made between metabolomics and genetic scores in the section "Comparative performance of metabolomics and genetics". It appears to me just a comparison of hazard ratios which is not equivalent to discrimination statistics or more complete formal assessment against clinical scores that follow. Please change title to "Comparison of hazard ratios for incident disease among -omic score subgroups"

The current title "Metabolomic scores from multiple timepoints" suggest more than two timepoints are being considered per subject. Please change to "Incident disease risk among participants with two metabolic scores"

• The four options for individuals with two metabolomic measures are 1) stay in high risk at both time points, 2) move form high to low, 3) move from low to high, 4) stay in low risk at both time points. I am not clear why the low to high score subjects are not shown in their own group. Are they included in the low-low group? Should they not be in their own group?
• The median/mean and range of days (or months/years) between the two measured for the 18,709 subjects should be reported in the results section.
• The sentence should be removed: "This level of stability reinforces the observation above that the scores provide information about disease risk many years into the future, but since the correlation is not perfect it also implies that multiple measurements can provide more information for risk prediction." I missed this detail on my previous review. I don't think one

can conclude this with certainty with the data presented.

Section Comparing multi-omics to existing clinical risk scores
…."the remaining four either have no difference, or inconsistent results between AUC and hazard ratio of the top decile". I assume for these four the hazard ratio is significant but the delta AUC is not. The delta AUC has much less power than the hazard ratio in a regression, so this is not unexpected (PMID 23296397).

Net reclassification metrics should only be reported stratified by cases and controls, and not combined (PMID 24240655).

The titles for Figures 1 and 3 should emphasize these are minimally adjusted models for age-sex and remove the term "performance".

Fig. 1. "Association between baseline minimally age and sex adjusted metabolomic score and 4-year incident disease in three national biobanks0".

Fig. 2. Ten-year disease risks and cumulative incidence stratified by minimally age and sex adjusted metabolomic score at baseline, polygenic score, and duration of follow up.

Fig. 3. Cumulative incidence of eight diseases stratified by age and sex minimally adjusted metabolomics risk score among subjects with two measures.

Fig 4. Predictive ability and calibration of models including clinical, polygenic and/or metabolomic scores

Please state more clearly which model is being tested for calibration in Fib 4b legend. For each disease, are these age/sex minimally adjusted models with metabolomics or models that includes clinical risk and PGS?

Table 2. If you are going to report a delta AUC with 3 significant figures, please also include a third decimal point measure (e.g. 0.75 should be changed to 0.752 or whatever the correct third digit is) for the AUC discrimination statistic (and the corresponding confidence intervals). Clarify whether age and sex are covariates included in the models: PGS, Metabolomics, PGS+Metabolomics.

Page 8, line 29. Modify FROM "We have shown that metabolomic scores can predict increased risk across" TO "We have shown that metabolomic scores can identify subjects at increased risk of disease across.."

Page 9, line 1-2. Modify FROM "Our direct comparison, and combination, of metabolomic and polygenic risk factorsdemonstrates the value of multi-omic scores" TO "Our direct comparison, and combination, of metabolomic and polygenic risk factors suggests value of multi-omic scores."

Please remove sentence "Combining these forms of information may also be useful in maximizing the predictive accuracy while avoiding the perceived determinism of fixed genetic predictions".

Change the sentence that follows FROM "Our data clearly demonstrates both that…" TO "Our data suggests that genetic factors contribute stable long-term predictive information and that these can be complemented, particularly….."

The term "cases" is used multiple times in the discussion section paragraph on page 9 between lines 29-4 (e.g. "In other cases"). I think it reads better if this is changed to "For some health traits," or "For some diseases," or "We identified some situations"….the word cases might be misconstrued for the actual cases of disease here and create confusion.

Please change FROM "we here defined those groups as the top 10% of each disease's risk score, but in practice different cut points will be appropriate for different diseases (e.g. in liver disease, risk is strongly concentrated in the top 1%). TO "we here defined those groups as the top 10% of each disease's risk score, but in practice different cut points may be appropriate for different diseases or cut off may not be necessary if the goal is to deliver improved prediction across all baseline risk strata". I suggest this above modification because I don't believe that the only way to implement these scores in clinical practice is by adopting an "actionable" cutoff although this is implied throughout this report. Both scores (PGS and metabolomics) have the potential to improve risk prediction across the full range of baseline risk and in many cases that will be useful on its own (especially when treatments exist that reduce risk no matter what the source of excess risk is – e.g. lifestyle and statins for myocardial infarction).

Reviewer #2

(Remarks to the Author)
I applaud the authors' efforts to incorporate reviewer comments/feedback and think that this revised submission is substantially improved.

My most substantive comment relates to the selection of clinical scores as benchmarks. Several of these scores are quite logical and well-justified: QRISK3, LLP, QDiabetes, QCancer. Other scores are screening tools for current diseases rather than prediction tools for future disease, and I'm not sure it makes sense to treat these in the same way – e.g., the PHQ-2 is a screen for current depression; what exactly is the interpretation when the PHQ-2 has an AUC for incident depression of 0.67

and PHQ-2 + metabolomics has an AUC of 0.68? At the very least, I would demarcate/group conditions where validated risk prediction scores exist (QRISK3, LLP, QDiabetes, QCancer) vs. others.

Minor: “This suggests that our scores are not forgetful: at any point in time both a person’s current metabolomic score value, and previous measurements that reflect their historic risk level, contribute information about risk of future disease onset.” Suggest rewording this sentence to be more objective and to sound less like ‘marketing’.

REVIEWER COMMENTS

We thank the reviewers and editor for their comments, based on which we have made substantial revisions to the paper, detailed below. To make it easier for the reviewers to see which sections have been most substantially rewritten in response to their comments, we have included some of the text in new sections in this point-by-point document, and uploaded a version with changes to the main text tracked, as well as a clean copy of the revised paper.

In addition, since this manuscript was submitted, we have completed metabolomic measurements for the full UK Biobank. We have thus rerun all analyses on an additional 217,000 samples and include these updated analyses in the paper. While most of the results are unchanged, increased sample size has:

- Made the meta-analysis of the Alzheimer's metabolomic score in Figure 1B now statistically significant, making this true of all 12 diseases.
- Shown the improved prediction of vascular dementia one year after baseline to be statistically significant.
- Enabled us to test performance in a wider combination of non-European ancestries for multiple diseases.

Reviewer #1 (Remarks to the Author):

I thank the authors for their response to my review regarding characteristics of excluded subjects from the biobanks, additional details re: LASSO models, and adjustment of statements related to proteomics. I am struggling a bit with the other aspects of the authors' response as it is collectively wrapped up in what is described as a 'philosophical difference' in objectives which are now no longer focused on demonstrating improvement in risk prediction over conventional approaches but rather some sort of 'proof of principle' that routine NMR testing can identify individuals at high risk of multiple disease and enhance a "national preventive health check program". The difficulty is the original intent seemed to have been indeed to make this claim (as per the original abstract and manuscript submission) and the revised manuscript continues to promote this idea even if the claims have been scaled back. Importantly, this message is being delivered by an investigative group that is dominated by employees of the company that sells the NMR test. Given these circumstances, I think it is critical to ensure that statements and conclusions made in this manuscript are unequivocally supported by the data and analyses. For this reason, I feel compelled to be more prescriptive in my critique this time around.

We thank the reviewer for this clear and honest appraisal. We have done our best to address their concerns, both by adding new analyses, and re-writing.

First, I recommend the authors further scale back and completely refrain from any commentary or speculation on whether the results of the study provide adequate evidence that NMR omics testing can be incorporated into public health practice to enhance national preventative check programs through screening testing performed on a regular basis.

We have removed the commentary and speculation the reviewer objected to. We have replaced the first paragraph of the paper, which previously discussed NHS Health check, with:

"Identifying individuals at elevated risk of future disease can help guide the use of preventative interventions. For example, in the UK the multi-variable QRISK score is used to identify individuals at high risk of cardiovascular disease who should adjust their lifestyle or begin taking cholesterol-lowering or blood pressure reducing medicine [1]. This concept of combining multiple measurements or risk factors into a single score has been extended to the use of 'omic data, such as polygenic scores (PGS) [2,3]. By adding up the contribution of different genetic variants associated with different diseases, PGS can identify individuals at elevated risk for multiple diseases [4] with one measurement (e.g. a GWAS array or genome sequence), and offer complementary information to traditional risk factors [5,6]. Metabolomic scores, based on adding up the contributions of multiple biomarkers measured from a blood sample, for example via nuclear magnetic resonance spectroscopy [7–9], have also been shown to predict many common diseases [10,11] including cardiovascular disease, type 2 diabetes [12], and all-cause mortality [13]. Furthermore, since the metabolomic scores may change in response to lifestyle and treatment (in contrast to PGS), they can also track changes in people's risk profiles. A few studies have suggested complementary value for genetics and metabolomics in cardiovascular disease and type 2 diabetes [14,15], but the combined use of these 'omics-based risk predictors has not yet been evaluated at scale."

We have also re-written the Discussion, including removal of any reference to health check programs or implementation in public health practice. The new Discussion reads:

"Discussion

We have shown that metabolomic scores can predict increased risk across a range of diseases, consistent with a previous report that analyzed a subset of the data described here[10]. We have replicated these findings in two additional national biobanks, and demonstrated consistent performance in three countries which have varying sample collection types (e.g. plasma vs. serum), enrolment criteria, fasting protocols, and electronic health record systems. As the biomarkers that constitute these scores are measured in absolute units, the scores can be computed without cohort-specific rescaling, which could aid in clinical translation of the scores. More than 1 in 4 individuals in these biobanks are in the high-risk group for at least one of the cardiovascular diseases, lung diseases, liver diseases or diabetes, where the high-risk group has at least 2.5-fold increase in risk (likely an underestimate of population levels, due to healthy volunteer bias).

Our direct comparison, and combination, of metabolomic and polygenic risk factors demonstrates the value of multi-omic scores. When good predictors exist from both 'omics data types (e.g. myocardial infarction, diabetes, colorectal cancer) the scores are complementary and together provide an improved combination of predictive accuracy and specificity. Combining these forms of information may also be useful in maximising the predictive accuracy while avoiding the perceived determinism of fixed genetic predictions: our data clearly demonstrates both that genetic factors contribute stable long-term predictive information and that these can be outweighed, particularly in the shorter term, by detectable differences in metabolomic risk

profile (driven, in part, by factors modifiable by lifestyle and treatment changes). Our analysis of follow-up samples underscores this: the scores show an intermediate level of stability five years apart, meaning they provide long-term risk prediction and can also track measurable change in risk in response to lifestyle changes or treatment. The observational nature of this dataset, as well potential survivor bias in the individuals who participated in the repeat sampling, limit our ability to make causal inferences about associations between changes in lifestyle factors and changes in scores. Future studies that explicitly measure metabolomics before and after an intervention are needed to fully explore how metabolic risk scores track changes in modifiable risk.

For 11 of the diseases we studied, we compared to existing scores used for risk prediction or screening in clinics. We identified clear cases where multi-omic scores alone outperformed classical clinical scores, including for myocardial infarction and liver diseases. In other cases the clinical scores perform better, often for understandable reasons. For instance, for type 2 diabetes the QDiabetes score we used includes Hb1Ac, the gold-standard diagnostic biomarker for the disease. In other cases, the clinical scores directly include vital causal risk factors, such as smoking for lung cancer. However, even in these cases the multi-omic score provides significant additional information: if we restrict to current smokers, the multi-omic score outperforms the clinical score for lung cancer (hazard ratio of 2.2 vs 1.8). Multi-omic scores provide statistically significant improvements when added on top of existing scores in all but one disease. Future work will be needed to quantify the extent to which these statistically significant improvements in prediction, when used to guide health interventions, could translate into improvements in population health, as has been recently studied in the specific case of cardiovascular disease[25]. In addition to accuracy, our results also allow us to test calibration across populations, which is vital to future applied use. While imperfect, the calibration of our scores is comparable to widely used tools like the pooled cohort equations for cardiovascular risk when compared, for example between the US and Canada [26].

Our study has several limitations which we could only partially mitigate. First, although biobanks are powerful for studying many diseases simultaneously, they typically have less detailed and well curated phenotypes than disease-specific clinical cohorts. For example, we use endpoint definitions based on broad categories of ICD-10 codes. In the case of cardiovascular disease, we have shown using more fine-grained coding does not substantially alter our conclusions, but clinically focused collections will help bring additional resolution. Furthermore, biobanks are known to have healthy volunteer bias, which can affect comparisons between these results and true rates of disease in the overall population. Second, we here defined those groups as the top 10% of each disease's risk score, but in practice different cut points will be appropriate for different diseases (e.g. in liver disease, risk is strongly concentrated in the top 1%). Third, these three biobanks are nearly all of European ancestry, and while we showed some promising results in the non-European ancestry subset of UK Biobank, analyses of more diverse cohorts will be essential. Finally, in comparing against clinical screening there are certain variables not measured in these biobanks that preclude a complete comparison, including of some widely used low-cost tests (e.g. fecal occult blood samples in colon cancer) and more intensive screening tools (e.g. CT screening for lung cancer or colonoscopies for colon cancer).

Many healthcare systems desire a more personalized and preventative model of disease management, including longitudinal monitoring of risk, early detection of disease and active, patient-centered management of risk factors. For this reason it will be valuable to understand how classical risk factors can be supplemented or replaced with newer predictors including metabolomics and genetics, and how and why these risk predictions change over time. We believe our results, and the extremely large dataset that underlies them, are an important part of this puzzle. Future work will need to focus on how we may incorporate additional new predictors to stratify disease risk further (e.g. recent work on proteomics has shown promise[27]), as well as how this next generation of prediction algorithms can be validated, made actionable and delivered the patients."

The analytic approach and the results arguably do not rise to that level of evidence for a several reasons (also echoed by reviewer 2): 1) the NMR scores have not been tested against the most advanced clinical scores/predictors and, for some of the diseases presented, in the correct clinical setting (see below) 2) the analyses do not correct for effects on scores that could be a consequence of the use of medications (while the authors present some compelling evidence for why statins do not artificially enhance discrimination for MI, they do not present results adjusted for statin use and do not systematically address a multitude of other drugs that could be preferentially used among cases or controls) 3) the statistical comparisons have been anchored on comparisons of hazard ratios rather than AUCs or other metrics of discrimination.

In addition to the re-writes mentioned above, we have also added analyses to address these points.

We replaced our comparison to a simplified non-omics score with a new section comparing to clinical scores that are recommend for use by relevant professional bodies. This section compares to these clinical scores using both AUC and hazard ratios, as the reviewer requests. Further details are included below in response to the reviewer's subsequent specific comment on this point.

We do not have data to systematically address the effects of a multitude of prescription drugs, but we have added one specific analysis on statins in response to Reviewer 3.

"To illustrate how the multi-omics score could augment the most widely used risk screening tool in the UK, we further focused on comparing QRISK to QRISK+multi-omics in individuals not using statins at baseline, which approximates the eligible group for a common use of these scores in prioritizing patients for statin treatment. The AUCs improve by 0.029 for myocardial infarction and 0.008 for stroke, which are equivalent to the whole-population values in **Table 2** (i.e. there is no interaction between these scores and statin usage). Considering net reclassification index, the continuous changes (i.e. net fraction of individuals whose risk score moves in the "right direction") are substantial: 0.53 for myocardial infarction and 0.19 for stroke.

Considered categorically (i.e. moving in the “right direction” between high and low risk groups) they are 0.07 and 0.02, respectively.”

We hope the Reviewer will agree that we have engaged substantively with their critique both by adding useful new results, and by changing how we present and interpret our previous results. We believe these combined changes address their concerns.

Along these lines, the authors themselves refer to a UK Biobank preprint in their discussion that tests the value of proteomics profiling in the UK Biobank (Gadd et al., reference 31 which now has a version 2 preprint). That manuscript adopts an approach to testing -omics versus clinical prediction that is in line with comments above and ideally should be emulated here. That manuscript also steers away from any discussion about implementing proteomics in clinical practice. Until a more comprehensive approach is adopted, I think the final sentence of the abstract should be removed or scaled back considerably.

We have removed the previous final sentence from the abstract.

In this context, I encourage the authors benchmark the metabolomics platform against more ‘complete’ or established clinical algorithms tailored to the disease whenever possible if emulating the approach adopted by the proteomics group is too much of a rewrite. These additional analyses should not be burdensome as they simply entail a rerun of scripts adding or substituting additional covariates to the base / clinical models in the same manner pack year smoking was considered. New data could then be added to the current figures and AUCs tables. For example, for diseases such as MI and ischemic stroke, the ‘clinical’ benchmark could be the QRISK3 score itself (restricted to the age range for the score) in the UK Biobank, or the inclusion of as many variables as available from the score in the baseline model. For diabetes, it could be the QDiabetes score.

We have now carried out a review of established clinical risk or screening scores in use in the clinic in the UK for these endpoints, and identified scores for 11 of our 12 endpoints (including the two mentioned by the reviewer). We now include analyses comparing our risk scores to these clinically-used risk scores, comparing both our score alone against the clinical score, as well as combining the two. We have presented these comparisons both in terms of hazard ratios for top deciles of the scores (Figure 4), and as AUCs and delta AUCs relative to the clinical model (Table 2). For some scores (including lung cancer and type 2 diabetes) the clinical scores outperformed our risk scores, but for all diseases but one, a combined score of clinical variables and ‘omics performs best, and in some cases (including myocardial infarction and liver outcomes) our multi-omic scores alone significantly outperformed tests used in clinic. We have described these comparisons even-handedly in the results (see below) and discussion (see above).

“Comparing multi-omics to existing clinical risk scores

We next compared our multi-omics predictions to published clinical risk scores both in terms of hazard ratios for the top decile and population-wide AUC. For all our diseases except Alzheimer’s we identified scores that are recommended for use either by the NHS in England and Wales or by professional bodies in the UK, EU or USA, and calculated them as accurately as possible using available variables in the UK Biobank (**Methods**). These scores vary in the types of variables they include (e.g. QRISK for cardiovascular disease includes several blood and anthropometric measurements, whereas PHQ2 for depression is based solely on two self-reported questions). The multi-omics scores perform significantly better for 10-year risk prediction of myocardial infarction, the two liver diseases and colon cancer, while the clinical scores perform significantly better for lung cancer, diabetes, and depression, and the remaining four either have no difference, or inconsistent results between AUC and hazard ratio of the top decile (**Fig. 4A, Table 2**). For all diseases except intracerebral hemorrhage (which is not well predicted beyond age and sex by any score we tested) a combined clinical+multi-omic score has significantly higher AUC than the clinical score, with increases ranging from 0.006 for lung cancer to 0.118 for alcoholic liver disease (**Table 2**). These results were also consistent for four years of follow-up (**Table S8, Fig. S12**).”

For COPD, lung cancer, and alcoholic liver disease, and cirrhosis, it would be helpful to add analyses in the figures and tables restricted to ever smokers and ever drinkers (some of this is in the text already).

We have included assessments of our scores in ever smokers vs never smokers (for COPD and lung cancer) and ever drinkers vs never drinkers (for ALD and cirrhosis), shown in Figure S6. This is presented in the new section of results, Investigations into the scores:

“Second, we sought to understand the extent to which our scores are driven by well-established behavioural risk factors for some of these diseases, in particular tobacco smoking and alcohol consumption. Both lung disease scores show attenuated, but still strong, association when conditioning on pack-years of smoking, suggesting that while they are partly driven by this behaviour, they capture additional information beyond the self-reported variables (**Fig. S6**). The performance of our lung cancer score is reduced to almost zero in never smokers, whereas our COPD score still has significant prediction in that group. Liver cirrhosis is equally well predicted across ever and never drinkers, and virtually unaffected by an adjustment for daily alcohol units. The adjusted alcoholic liver disease prediction is somewhat reduced but remains very strong (**Fig. S6**).”

The discussion section is not optimally written. The first paragraph section about AUCs and clinical risk scores is confusing. The first few sentences of the final paragraph re: incorporating NMR metabolomics in public health practice are premature and should be avoided for reasons described above. Lastly, a comprehensive paragraph describing the limitations to these analyses

is missing (again, see Gadd et al. for a good example to emulate). For several diseases presented, it simply is not possible to test against the current 'gold standard' approach to preventing or identifying specific disease early. For example, comparing NMR testing vs. screening with a CT scan for lung cancers among smokers is likely not possible to do with the data at hand. The same goes for 1) colon cancer and fecal occult blood tests and/or colonoscopies, 2) more sophisticated screening questionnaires for unhealthy alcohol use (e.g. AUDIT/AUDIT-C, CAGE), 3) COPD and pulmonary function testing among smokers. The discussion section needs to be more explicit about these limitations of the analytic approach presented (and the data available).

We have rewritten the discussion (see text pasted above), removing the commentary on implementation. We have added a paragraph on limitations, as suggested (see text pasted above).

Re: alcohol liver disease and cirrhosis, is it possible that the NMR testing is generating a signature of active heavy drinking? If yes, this should be explicitly mentioned in the manuscript as this may have major implications when trying to implement this test, some of these ethical.

While we think it is highly likely that the ALD score is picking up metabolomic signatures related to long-term alcohol consumption, we do not believe the scores are capturing active heavy drinking in the way the reviewer implies (i.e. that it is a hidden or proxy test for drinking right now). For example, the correlation between the alcoholic liver disease score and daily alcohol unit consumption is 0.18 and the liver cirrhosis score is virtually unaffected by an adjustment for alcohol consumption. Rather, these scores are capturing the downstream effects after prolonged exposure, much like the diabetes score is likely affected by obesity, lung disease scores by smoking, and heart disease scores by lifestyle.

Table 2 is a nice addition of AUC but can be improved. I am not clear if the first three columns have any clinical variables. I think they are PGS and metabolomics or combined alone but it would be helpful to make that clear in the table as a footnote. Clinical should probably be presented as first column. Some diseases should have more than one entry presenting comparisons to more complex models as suggested above. Also, why not include clinical+PGS, clinical + omics/PGS?

We have updated Table 2 as suggested. We also now include delta AUCs, relative to the disease-specific clinical risk models, for the 11 diseases where one is available.

Reviewer #2 (Remarks to the Author):

Thanks for the opportunity to review this manuscript. The authors test the predictive utility of 36 metabolic biomarkers assayed using the Nightingale platform in ~477,000 individuals across three cohorts. The authors further compare risk stratification by metabolomic scores vs. polygenic risk and, as expected, observe greater hazards associated with high metabolomic vs. high polygenic risk. Incorporating metabolomics appears to improve prediction to varying extents across disease outcomes. Overall, the analysis has several strengths and suggests

potential clinical utility of these biomarkers in practice. Although I agree with the authors that each disease needs its own careful consideration in future analyses, and although the authors appear to be de-emphasizing improvement in risk prediction in this revised submission, I think the manuscript would benefit from a bit more demonstration of incremental utility and new insights compared with previously published work examining Nightingale metabolomics in the UK Biobank and elsewhere, especially given the 'biomarkers-for-all' framing of Discussion.

We thank the reviewer for their comments. We have added more information on incremental utility compared to existing scores (see details above), and highlighted differences compared to previous work, such as the combined analysis of metabolomic and genomic scores, inferences from two time points, and the stability of predictions over time. As mentioned in our response to Reviewer 1, we have also de-emphasised the "biomarkers-for-all" message, and tried to present all the comparisons even-handedly.

As noted above, although I agree with the authors that each disease needs its own careful consideration and that doing this for all diseases may be beyond the scope of this paper, it would be helpful and informative to 'dig into' one or a couple of use cases where robust risk prediction approaches are developed and already in clinical practice. For example, as noted by the authors, the QRISK and FINRISK scores are validated for ASCVD outcomes (coronary heart disease events and ischemic stroke). I suggest the authors provide delta AUC and net reclassification indices for incident myocardial infarction and ischemic stroke when applying the metabolomic score on top of QRISK/FINRISK to statin-untreated patients (i.e., the populations where these scores are used to inform statin prescribing decisions); this would reasonably simultaneously address several of the comments raised by previous Reviewer 2 in response to the prior submission.

We have added a more extensive comparison to clinically-used risk scores (described in our response to Reviewer 1 above. In addition, we have now carried out the comparison between QRISK3 and our multi-omic scores that the reviewer suggests, and presented them in the text. The delta-AUCs are significant for both MI and stroke in non-statin users, though larger in the former case (0.03) than the latter (0.01). The NRIs are similarly better in MI than stroke, though again both show improvement.

Table 2: It would be helpful for the authors to provide AUCs for clinical score + PGS and, separately, for the clinical + metabolomic score (in addition to clinic + combined omics), as well as P-values for comparisons of each approach to the clinical-only score (e.g., using the Delong-Delong test).

We have made these changes to Table 2.

Minor:

Unless I am missing it somewhere, the actual metabolites used in the score for each disease

outcome are provided only in Table S3, which makes it difficult to ascertain how the scores are similar/different across conditions. Can the authors provide this info in a separate, easier-to-read supplementary table?

While, as the reviewer says, this data was present in our previous version, we agree that it was not presented in an easily digestible form. We have made an additional figure using the data from Table S3 that makes it easier to ascertain the similarities and differences between the scores (Figure S3).

The authors should add P-values for deltaAUC to Table S5.

We have added these.

It appears there may be an error in Figure S4b – the forest plot appears duplicated, whereas the headings "Myocardial infarction" and "Ischemic heart disease" each apply only to a subset of the ICD codes currently listed under both headings.

We acknowledge that the previous version of Figure S4 was unclear. We have updated it to convey the intended message more clearly.

The authors might consider highlighting for the reader who is more peripherally familiar with the multi-omic biomarker literature (e.g., in the Discussion) how the present analysis is distinct from/additive to the Buergel et al., Nat Med 2022 paper.

We have added information on this topic in the Discussion.

Reviewer #3 (Remarks to the Author):

In this paper the authors trained risk prediction scores from NMR based metabolomics features and polygenic risk scores for 12 diseases selected based on these being leading causes of disability-adjusted life years in high-income countries. In half of the UKB data Cox proportional hazards models were applied to predict incidence of the 12 diseases based on feature selection by Lasso (and five-fold cross-validation) from 36 clinically validated blood based metabolomics features.

Using the weights estimated in UKB, the other half of UKB and the Finnish and Estonian biobanks were used to test the relative and joint performances of the metabolomic scores and PGSses for predicting the 4 years incidence of 12 diseases, and 10 years prediction across follow up compared to clinical markers as well as the 10 years performance of repeated metabolomic measures at different time points.

This is an very interesting study of large scale metabolomics and genetics data in three large national biobanks. Many studies aim to generate risk predictors from omics data and the strength of the current study is the large scale analysis allowing a comparison to genetic risk factors (PRS) and clinical risk factors in three studies. The high-risk groups showed consistently increased risks in all biobanks (3 countries, across plasma vs. serum sample types, various enrolment criteria, and fasting protocols).

We thank the reviewer for their positive comments.

The ultimate goal of the study is not clearly formulated. It does not seem to be to test for the optimal predictive information for these DALYS in the 249 metabolomics features as compared to specific clinical risk markers for these diseases, which is a shame given the large dataset. The focus seems more on whether 36 clinically validated features (out of 249) can ultimately be used in real-world public health settings. Authors should clarify the goals, because in a dataset as large as the one described here one would expect a search for the optimal prediction from available features as well, or if focused on 'real-world public health settings' on meeting or improving the gold standard. But what is the standard for clinical risk estimation in the clinic for these DALYs, the only hint is the last paragraph indicating a comparison to NHS health check. Is that the ultimate aim ?

As described above, we have amended how we have described our results in response to reviewer comments. We no longer make an explicit case for real-world implementation of omic scores. Instead, we focus on the comparisons between metabolomic and genomic scores, the observations from two time points, and a comparison (without explicit recommendation) to standard scores used in clinic for these diseases.

As suggested, we have trained models using all 249 biomarkers, and shown results in a new Supplementary Figure S5. For some diseases (e.g. T2D & COPD) this significantly improves prediction, though in other diseases there is little change, presumably because of the correlation structure of the measured biomarkers. The new section reads:

"First, we assessed the performance of scores using all 249 metabolomic biomarkers we measured, rather than the 36 clinically validated biomarkers described above. Only diabetes and COPD showed consistently improved performance using the extended metabolomics (**Fig. S5**), likely because many of the biomarkers are correlated, and our clinically validated subset captures a large fraction of the total available information in most cases. In some cases, such as Alzheimer's and the liver diseases in the Estonian biobank, the extended metabolomic scores do not replicate as well as the simpler scores. Taken together, these results suggest some diseases may benefit from additional molecular measurements, but care must be taken that they do not capture cohort-specific effects which are less transferrable."

Even though we now do not discuss implementation explicitly in this paper (per the requests from other reviewers), we continue to focus on the clinically approved 36 biomarkers that, for example, we are currently using in clinical practice in Finland.

The clinical score is generic, based on age, sex, smoking pack-years, BMI, systolic blood pressure, HDL cholesterol, total cholesterol, and weekly alcoholic drinks. Are these the variables used in the clinic for all DALYs ? Justification and reasoning on the clinical scores and variables used for comparison of prediction of these 12 DALYs must be better presented. One can imagine that missing from the clinical variables is CRP or albumin, also regularly used in the clinic, and also in commercial health checks; and certainly present in the cohort studies, but why then not included here for comparison.

We have now included comparisons to currently recommended screening tools for our 12 diseases. Notably, none of these use albumin or CRP, indicating that these blood measurements, despite being very widely used, are not routinely incorporated into screening. We have noted this in the discussion.

Another point is : there is comparison to clinical prediction across 10 years trajectories but not across the 4 years that the paper starts with, would that not be also relevant for public health ?

We have added a Supplementary Table on clinical prediction for 4 years of follow up, which is similar to the 10-year results.

The result section does not summarize much on differences between the cohorts (except collection) and yet in Figure 4 they show different results here and there so do discuss Table 1/S2 a little better. (differences in health in the three studies, which may be relevant for discussion Fig 4).

We have added discussion of these items, as suggested.

As indicated it can be more clear from the start of results that all scores were based on 36 clinically validated markers out of 249 available ones measured in total.
It must be shown in supplements which features are most prominent in the various scores. One can not see which features are most informative, which seems especially interesting for the four conditions for which metabolomics adds to clinical risk prediction. And also to understand why scores for COPD, Lung cancer and depressive disorders cluster as can be seen in Figure S6.

We have added a new figure that visualizes the weights for each biomarker in each of the metabolomic disease scores (Figure S3). On the specific comment related to lung diseases and depression, this is likely partly due to smoking (Figure S6).

The goal of future rapid translation seems a bit in the way of trying to get the best predictions which should ideally have been part of this epidemiological study, is the maximum for this

platform reached when using the 36 features because the rest is redundant or have studies shown that the 36 only get halfway as compared to the unused features ? It is a logical analysis to explore to what extend other features could have improved prediction. In other metabolomics work usually about 65 independent features are used for feature selection, the choice for the 36 is based on allowing standardized future application of the score, not to test the most optimal prediction. But which metabolites are missing if you compare the 36 to the metabolites strongly associating to outcomes in the papers referred to about previous work (15-18), do you expect to gain predictive power if these were added in future.

We fit additional models with the complete set of 249 biomarkers and tested the scores in the UK Biobank, Estonian Biobank and THL Biobank (Figure S5). Including all the biomarkers leads to slight improvements in some of the risk models, but for most diseases has inconsistent or no effects across the different biobanks. This result can largely be explained by high correlations between many of the biomarkers.

Table 2 is highly relevant when considering the translational value of the scores and platform.

We agree, and have added more detail in the table (including delta AUCs compared to clinically-used model) and in the text to emphasize this data.

Some observations can be phrased more clearly I think

1. Considering the 10 years risk prediction, omics scores based on the 36 features (27 separate features and 9 ratios) used here, only exceed the clinical prediction above marginal for T2D, MI, and the liver conditions. And that clinical prediction does not even include albumin or CRP. Leaves the reader with the feeling: could prediction have been increased when more measured features were used. Table 2 shows that for 7 out of 12 diseases, the prediction improvement as compared to the clinical variables is marginal (.01 increased AUC). For Colon cancer this is slightly higher .02) and for MI, T2D, and the liver prediction is really improved.
2. Clear is that genetic scores hardly contribute to the improvement that was observed. In the four improved cases it is mainly (MI) or only (T2D and liver diseases) metabolomics that increases prediction..
3. From a translational point of view it is highly valuable that the variable weights estimated in the UKBB could be used in other studies without normalizing the biomarkers within each study separately.
4. So one may conclude that metabolomics alone based on the clinically validate features generates the same information as the clinical score for most DALYs and improved it in four conditions
5. But then again it is critical what is currently used and would clinic and public health agree on the clinical variables selected here.

We thank the reviewer for summarizing these points, which we agree are interesting and valuable, and have rewritten our discussion to further emphasize.

It is clear that the metabolomics scores from the 36 features selected are not specific for these DALYs, but neither would the clinical score be considering the selected variables. Please include the extent to which the clinical features relate to the DALYS Since the predictors are not very specific, how can they be translated, were the clinical predictors less correlated ?

We have now added the disease-specific clinical scores to Figure S9 on score correlations.

The analysis of repeated measures is very interesting indeed; could authors discuss whether there may be impact of excluding disease events between baseline and repeat visits but leaving the persons in. Persons could be missing in the repeat measure because they were unhealthy at baseline, had their event and did not pay the repeat visit or died; these would be left out. So those surviving the event and paying the next visit are in the second visit data and then may have no event anymore unless a recurrent one. So, for those who had an event maybe treatment lowered their metabolomic score.

Our analysis of the repeat samples only included individuals that did not have an event by the time of the repeat sample was taken, so individuals who had an event between the baseline and repeat visits were excluded from assessment. As the reviewer points out, this and other features (not having had an event, not having died, not dropping out of the study) makes the repeat samples a non-random selection of the cohort. We now discuss this in the paper.

The abstract mentions only that lifestyle could have changed the score but what was investigated only explained a small portion of the change. The sentence ‘ People whose scores change have different future risk of disease, suggesting that repeat measurements capture the benefits of lifestyle change’ would have to be rephrased in something like: repeat measures capture changes bot health statis and future disease risk possibly due to treatment, lifestyle changes or other beneficial factors.

We have rephrased this sentence according to the reviewer’s suggestion.

The discussion touches proteomics and age of the study populations very briefly and no mentioning of how the metabolomics scores may be improved by the other available metabolites ot other metabolomics technologies also used in epidemiological studies (Mass spec based lipidomics, Metabolon platform). The proteome was studied in UKBB , so do authors foresee specific proteins that may add to predictions ? What are expectations.

We agree this is an interesting point, but as we have been asked not to speculate beyond our data by Reviewer 1 we have not addressed it in the paper.

Because participants are still young (also for relating to frailty by the way) , perhaps the predictions improve in older persons where generic scores representing health status can be useful not only for lifestyle improvement.

This is an interesting point, though again we have not mentioned it explicitly here.

Overall the paper is very interesting by its large scale, diversity of cohorts, comparison of two omics PRS and showing that the validated features included in the score are robustly indicating increased risks in all biobanks (3 countries, across plasma vs. serum sample types, various enrolment criteria, and fasting protocols) holding promise for using metabolomics as generic health check to support preventive health care in the population at large. It can be more clear in showing whether and how predictions may be improved, and perhaps a more realistic comparison with clinical variables as they are measured in real life situations for these DALYs in clinic or population. Your study has to be convincing that once blood sampling would be included in population wide health checks, the metabolomics will exceed prediction over clinical variables that would then include common blood based markers.
It should also be pressed that it is of great interest to the biomarker field to explore if one can turn affordable, robust omics biomarker assays into practically usable tests and the paper is a serious effort in testing this out.

We thank the reviewer for their kind comments, and their valuable feedback on our study.

Reviewer #1 (Remarks to the Author):

I would like to thank the authors for their modifications and for including additional data. It's quite remarkable that the entire UK biobank has now been profiles with the metabolomic platform. I think overall the manuscript has been strengthened substantially.

We thank the reviewer for these positive comments.

However, I continue to feel the need to be prescriptive to manage potential conflict of interest. Thus, I have these very specific suggestions to make one last time. A vast majority of these are simple changes to the text but a couple may require new simple analyses to clarify.

We have made changes to the manuscript in response to these concerns, detailed below.

ABSTRACT: The abstract highlights positive results for all comparisons that are minimally adjusted but does not summarize the results of the critical last set of analyses comparing clinical scores to clinical + metabolomic scores. I may suggest replacing the current final sentence of the abstract with the following adjustment: "Lastly, we assessed the incremental predictive value of metabolomic scores over existing clinical risk scores for multiple diseases and found modest improvements in discrimination for several diseases whose clinical utility, while promising, remains to be determined." I believe this summary of findings is extensively described in the discussion section already but is noticeably absent in the abstract.

We have made the suggested change.

TEXT: The term "future risk" and "future disease onset" is used throughout the manuscript. I do not think the word "future" is needed. It is redundant.

We have deleted the word "future", which we agree is more succinct.

I am somewhat dissatisfied with multiple section headings and legend titles to the figures. They are either not descriptive enough or suggesting something more than what is being delivered in that section. In multiple section headings, the term "performance" or "risk prediction performance" is used, when performance is not being formally assessed at least not against the 'gold standard'. Below I make suggestions for changes in titles of sections and figures. Within some sections I also make additional suggestions for change.

Please change section title "Risk prediction performance" (Page 3, line 25) to "Baseline age and sex minimally adjusted metabolomic scores and incident disease"

We have made this change.

I would change the next title from "Investigations into the scores" to "Sensitivity and

Subgroup Analyses of minimally adjusted scores”

We have made this change.

We note that the reviewer has proposed the phrase “minimally adjusted” in several of their comments. We agree that it is useful in the headings, to make it clear that those sections do not cover comparisons to scores with more clinical variables, and so have changed it as described above. The reviewer has also consistently requested precise and succinct language (which we have worked hard to align with), and we believe that following this approach, it is best to refer in the text and figure legends to the adjustments by neutrally stating what they are (e.g. age and sex), rather than use this phrase repeatedly.

There is no formal comparison of performance being made between metabolomics and genetic scores in the section “Comparative performance of metabolomics and genetics”. It appears to me just a comparison of hazard ratios which is not equivalent to discrimination statistics or more complete formal assessment against clinical scores that follow. Please change title to “Comparison of hazard ratios for incident disease among -omic score subgroups”

We have changed this to “Comparison of hazard ratios among metabolomic, genetic, and combined scores”, as we think this is clearer “-omic score subgroups”.

The current title “Metabolomic scores from multiple timepoints” suggest more than two timepoints are being considered per subject. Please change to “Incident disease risk among participants with two metabolic scores”

We have made this change.

• The four options for individuals with two metabolomic measures are 1) stay in high risk at both time points, 2) move form high to low, 3) move from low to high, 4) stay in low risk at both time points. I am not clear why the low to high score subjects are not shown in their own group. Are they included in the low-low group? Should they not be in their own group?

In the past version these were included in the low group, as we were mainly interested in the trajectory of individuals who began in the high-risk group. But we agree that the low-to-high group is also of interest, so we have now updated Figure 3 to split the survival curves out into four groups, as suggested.

• The median/mean and range of days (or months/years) between the two measured for the 18,709 subjects should be reported in the results section.

We have added the median, mean and range of the time difference, in years, between the two measurements to the results text:

“who returned for a repeat visit approximately four and a half years after they initially enrolled in the study (median time difference 4.4 years, mean 4.3 years, range 2.1-6.9 years).”

• The sentence should be removed: “This level of stability reinforces the observation above that the scores provide information about disease risk many years into the future, but since the correlation is not perfect it also implies that multiple measurements can provide more information for risk prediction.” I missed this detail on my previous review. I don’t think one can conclude this with certainty with the data presented.

We have deleted this sentence.

Section Comparing multi-omics to existing clinical risk scores
….”the remaining four either have no difference, or inconsistent results between AUC and hazard ratio of the top decile”. I assume for these four the hazard ratio is significant but the delta AUC is not. The delta AUC has much less power than the hazard ratio in a regression, so this is not unexpected (PMID 23296397).

In one case this is true, in two cases neither metric shows a significant difference, and in the fourth (COPD), the delta AUC is significant, but the decile HRs are ~identical. This could be because the scores perform differentially at assessing risk in the bottom 90% of the population (relevant to the point the reviewer makes below about one line in the discussion).

Net reclassification metrics should only be reported stratified by cases and controls, and not combined (PMID 24240655).

We have replaced the NRI values in the text by values stratified by cases and controls:

“Considering net reclassification index, the continuous changes (i.e. net fraction of individuals whose risk score moves in the “right direction”) are substantial: 0.21
t
h
e
The titles for Figures 1 and 3 should emphasize these are minimally adjusted models for age-sex and remove the term “performance”.
r
Fig. 1. “Association between baseline minimally age and sex adjusted metabolomic score and 4-year incident disease in three national biobanks0”.
h
We agree “Association between…” is better than “performance”. However, both the details on what is adjusted for, and the follow-up time are specified in the legend. It is common practice to have concise figure titles followed by detailed legends. For
i
example, the Gadd *et al* protein score paper (previously highlighted by this reviewer
a
as an example to emulate, now published in *Nature Ageing*) includes general figure
t
titles such as, “Value offered by ProteinScores for incident outcomes in the UK
B
Biobank.” and “Exploration of the type 2 diabetes ProteinScore.”
t
We have therefore changed this to “Association between metabolomic scores and disease onset in three national biobanks.”
n
Fig. 2. Ten-year disease risks and cumulative incidence stratified by minimally age
”

and sex adjusted metabolomic score at baseline, polygenic score, and duration of follow up.

We have changed this to, “Associations between metabolomic and polygenic scores and disease onset.” For this three-panel figure, we think it is difficult to capture all the elements in the title, and it is better to put the details in the legend, as described above.

Fig. 3. Cumulative incidence of eight diseases stratified by age and sex minimally adjusted metabolomics risk score among subjects with two measures.

We have changed this to, “Risk of disease onset for eight diseases stratified by metabolomics scores at two time points.”

Fig 4. Predictive ability and calibration of models including clinical, polygenic and/or metabolomic scores

We have made this change, and agree this title needed updating in light of the changes to the rest of the text in the previous revision.

Please state more clearly which model is being tested for calibration in Fib 4b legend. For each disease, are these age/sex minimally adjusted models with metabolomics or models that includes clinical risk and PGS?

We have clarified this to, “by 10 equally sized deciles of absolute risk predicted by a metabolomic score adjusted for age and sex”.

Table 2. If you are going to report a delta AUC with 3 significant figures, please also include a third decimal point measure (e.g. 0.75 should be changed to 0.752 or whatever the correct third digit is) for the AUC discrimination statistic (and the corresponding confidence intervals). Clarify whether age and sex are covariates included in the models: PGS, Metabolomics, PGS+Metabolomics.

We have rounded the delta AUC’s to the nearest percentage point, as for the most part the confidence intervals show that the third decimal point of AUC is not estimated with precision. We have shown values below 0.01 to the nearest 0.001 for clarity, as some of these small differences in the more common diseases (e.g. T2D) are still stastically significant. We believe the Table now gives the reader an accurate understanding of these comparisons.

Page 8, line 29. Modify FROM “We have shown that metabolomic scores can predict increased risk across” TO “We have shown that metabolomic scores can identify subjects at increased risk of disease across..”

We have changed to “We have shown that metabolomic scores can identify individuals at increased risk across…”

Page 9, line 1-2. Modify FROM “Our direct comparison, and combination, of metabolomic and polygenic risk factorsdemonstrates the value of multi-omic scores”

TO “Our direct comparison, and combination, of metabolomic and polygenic risk factors suggests value of multi-omic scores.”

We have made this change.

Please remove sentence “Combining these forms of information may also be useful in maximizing the predictive accuracy while avoiding the perceived determinism of fixed genetic predictions”.

This sentence uses careful language (“may also be useful”), and raises a point that is a topic of debate in the field, e.g. PMID 35354965, which notes, “Health-care providers and patients may perceive polygenic risk as deterministic despite evidence that such risk can be mitigated by lifestyle changes as well as drug therapy”. We think it is appropriate for the Discussion section of our paper. We have added the reference noted above.

Change the sentence that follows FROM “Our data clearly demonstrates both that…” TO “Our data suggests that genetic factors contribute stable long-term predictive information and that these can be complemented, particularly…..”

We have made this change.

The term “cases” is used multiple times in the discussion section paragraph on page 9 between lines 29-4 (e.g. “In other cases”). I think it reads better if this is changed to “For some health traits,” or “For some diseases,” or “We identified some situations”….the word cases might be misconstrued for the actual cases of disease here and create confusion.

We agree this was confusing. That section now reads:

“
restrict to current smokers, the multi-omic score outperforms the clinical score for lung cancer (hazard ratio of 2.2 vs 1.8).”

Please change FROM “we here defined those groups as the top 10% of each disease’s risk score, but in practice different cut points will be appropriate for different diseases (e.g. in liver disease, risk is strongly concentrated in the top 1%). TO “we here defined those groups as the top 10% of each disease’s risk score, but in practice different cut points may be appropriate for different diseases or cut off may not be necessary if the goal is to deliver improved prediction across all baseline risk strata”. I suggest this above modification because I don’t believe that the only way to implement these scores in clinical practice is by adopting an “actionable” cutoff although this is implied throughout this report. Both scores (PGS and metabolomics) have the potential to improve risk prediction across the full range of baseline risk and in many cases that will be useful on its own (especially when treatments exist that reduce risk no matter what the source of excess risk is – e.g. lifestyle and statins for myocardial infarction).
We agree, and have re-written that section as:

“
ase’s risk score, but in practice different cut points may be appropriate for different

range of baseline risk in the population."

Reviewer #2 (Remarks to the Author):

I applaud the authors' efforts to incorporate reviewer comments/feedback and think that this revised submission is substantially improved.

We appreciate the reviewer's kind feedback.

My most substantive comment relates to the selection of clinical scores as benchmarks. Several of these scores are quite logical and well-justified: QRISK3, LLP, QDiabetes, QCancer. Other scores are screening tools for current diseases rather than prediction tools for future disease, and I'm not sure it makes sense to treat these in the same way – e.g., the PHQ-2 is a screen for current depression; what exactly is the interpretation when the PHQ-2 has an AUC for incident depression of 0.67 and PHQ-2 + metabolomics has an AUC of 0.68? At the very least, I would demarcate/group conditions where validated risk prediction scores exist (QRISK3, LLP, QDiabetes, QCancer) vs. others.

We agree that this is a good point. When adding these comparisons we did consider restricting ourselves, for example, to the four listed above which best match our analysis. However, we felt including approximately relevant scores for as many diseases as possible was the best way to respond to the previous round of reviews.

We have added to the discussion:

"Furthermore, it will be important to consider whether and how multi-omic data fit in the diverse clinical contexts represented by these scores. QRISK3 and QDiabetes are mainly tools for identifying high risk individuals who would benefit from primary prevention efforts, COPD-PS, PHQ2, AUDIT and FLI are screening tools for prioritising further investigation in individuals who may have undiagnosed disease, and QCancer and LLP are used for both primary prevention and to aid early detection."

Minor: "This suggests that our scores are not forgetful: at any point in time both a person's current metabolomic score value, and previous measurements that reflect their historic risk level, contribute information about risk of future disease onset." Suggest rewording this sentence to be more objective and to sound less like 'marketing'.

We have changed this sentence to read, "This suggests that both a person's current metabolomic score value, as well as previously measured score values, contribute information about risk of disease onset."