nature portfolio

Peer Review File

Gut Microbiome Wellness Index 2 Enhances Health Status Prediction from Gut Microbiome Taxonomic Profiles

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this study, the authors introduced GMWI2 as a novel predictor of human health. GMWI2 was calculated as a weighted sum based on the presence/absence of microbial taxa identified from stool shotgun metagenomic data. The weights were determined through a LASSO-panelized logistic regression model using a pooled dataset comprising 8069 stool shotgun samples from existing literature. The predictive performance of the GMWI2 classification model was assessed on both the initial discovery dataset of 8069 samples and four additional external datasets.

Overall, this is a well designed study. One major technical concern is the potential risk of data leakage during model testing. The various validation procedures described in the section "Enhanced classification of healthy and non-healthy gut microbiomes with GMWI2" and the section "Evaluating the robustness of GMWI2 across different study populations" all employed a subset of the initial 8069 samples as test sets. However, the entire set of 8069 samples was used to compute the logistic regression coefficients and GMWI2 scores. Consequently, the GMWI2 scores inherently contain information from the test sets, irrespective of the chosen validation strategy. To mitigate the issue, the logistic regression coefficients and cutoff parameters should be estimated exclusively on the training data. While the authors did evaluate the GMWI2 model's performance on four external datasets, these evaluations did not specificially address the model's ability to distinguish between healthy and non-healthy samples.

Another evidence of potential data leakage can be observed in the sentences discussing the selection of the regularization parameter C: "The selection of the regularization parameter C was achieved through hyperparameter tuning, where we evaluated various candidates and selected the value (C = 0.03) that yielded the optimum classification performance in inter-study validation (ISV)". The parameter C and other hyperparameters should be evaluated exclusively on the training set, along with logistic regression coefficients, for each given train-test split. The reestimation of these parameters helps ensure the robustness and integrity of the model's performance assessment.

In addition, the authors used balanced accuracy (averaged proportion of correctly classified healthy and non-healthy samples) to evaluate the performance of the GMWI2 classification model. However, the model can output a third category, "not determined"(or defer), in addition to classifying samples as healthy or non-healthy. To gain a comprehensive understanding of the model's performance, it would be essential to know, for each validation, the number or percentage of samples classified as "not determined". In extreme cases, it is possible for the model to achieve 100% balanced accuracy while having limited practical value. For example, the model may correctly predict 1 healthy sample and 1 unhealthy sample but unable to determine the label for the remaining 98 samples.

I also have some personal comments on the significance of the study. The authors claimed that the model "pave the way for the early screening of adverse gut health shifts" and is "adaptable for researchers interested in the translational applications of human gut microbiome science". While these claims hold considerable promise, I encountered some difficulty in understanding the practical utility of the model in translational applications. If an individual uses this tool and the output indicates an "unhealthy" status, the interpretation becomes a pivotal question. Does an "unhealthy" result indicate that the individual is currently experiencing health issues or that the individual may develop one of the 12 diseases you studied in the future? Furthermore, what actions should the individual take in response? From a practical standpoint, if the person is already displaying symptoms, the microbiome test may not provide additional value in confirming their health status. On the other hand, if the individual is asymptomatic, he/she may have to wait until symptoms manifest before taking any proactive measures. This implies that merely categorizing individuals as healthy or unhealthy may not offer actionable guidance.

Finally, I want to express my reservation regarding the following statement "Any additional information required to reanalyze the data reported in this paper is available from the corresponding author upon reasonable request." In my opinion, this statement may not be acceptable. At the minimum, the microbial taxonomic profiles of the 8069 samples should be made available for download. Without access to these essential data, the reproducibility of the results presented in this manuscript cannot be rigorously accessed.

Minor comments:

1. Line number is missing. Page and line numbers would make it a lot easier for me to provide feedback.”

2. "…long-term safety and efficacy of FMT in treating patients with chronic diseases, computational tools like GMWI2 could be useful in assisting with the selection of healthy donors and stool samples." How does GMWI2 address the safety issue? The taxa with negative coefficients do not include many invasive pathogens such as Enterococcus faecium/faecalis.

3. It is unclear how the cutoff parameter c was determined for each validation strategy. Is there a single, universal c derived from the 8069 samples and applied uniformly across all scenarios, or is this parameter reestimated for each train-test split? In addition, the values of c need to be reported.

Reviewer #2 (Remarks to the Author):

Understanding the gut microbiome and its role in disease represents an ongoing challenge for the research field. The ability to rapidly and efficiently assess a participants microbiome as health associated or diseased has the potential to provide key insights and understandings in many contexts. Too often statements such as dysbiotic are used without clear definition or quantification limiting comparison between individuals, cohorts and studies more broadly. The authors present an refined version of their previously published index, the Gut Microbiome Wellness Index (GMWI2) to address this challenge.

The presented work represents an advance in standardising reporting of microbiome composition in stool samples and linking this to general patient state; however, there are some key limitations both in terms of the overall approach, the specific implementation and the limits in comparison. Most notably the comparisons being made to only existing GMWI and traditional diversity measures rather than current best practice analysis methods. Very few high quality studies, including many of those used as the exemplars rely purely on diversity to determine biological relevance of the findings. These additional comparisons should be included and this situation more clearly articulated in the manuscript to ensure appropriate application of the developed index. Further specific considerations are included below:

1. The choice of samples for inclusion and testing is of critical importance to the model overall. How does the definition of healthy impact the overall accuracy and robustness of the model and which factors defining health (e.g. BMI, age, etc.) contribute most strongly to the prediction.

2. While small size studies were reasonably excluded from the training set, however, is there an observable difference when evaluating robustness of the model of these samples. If this model is to be broadly applied it is essential to demonstrate it is equally applicable on small datasets more typical of those it is likely to be applied to. Further assessments of these types of datasets would be highly valuable in demonstrating usefulness of the approach

3. It is notable that multiple taxonomic levels were identified with non-zero coefficients. Given the dependence of the taxonomic hierarchy what impact does this have on the model?

4. While the data is presented comparing the GMWI2 to more basic measures (e.g. diversity) using the training dataset, further comparisons on independent sample sets to more accurately define where the application of the GMWI2 model provides maximum value would be highly advantageous.

5. The authors highlight the ability of incorporating magnitude of the GMWI2 scores to improve accuracy, biologically this is a highly rational approach, however, further exploration of the optimal values would be of substantial interest. While such values would

be impacted by the nature of the training sets, etc. some attempt should be made to achieve this.

6. The authors should be commended on dedicating time to evaluating the impact of individual studies on the robustness of the GMWI2 approach. It is interesting to note the ISV classification accuracy does not appear to correlate with any of the parameters shown. Does this indicate the quality of the definition of health may be a major contributor to the observed results. The authors should include further analysis to explore this phenomenon.

7. The authors present a small number of case studies to suggest the suitability of the GMWI2 approach. These should be subjected to far greater analysis and specific evidence to demonstrate both the applicability of the GMWI2 approach and the generalisability of the claims of suitability for application to that disease state if they are to be included as evidence to support the suitability of the method in these contexts. For example, the authors suggest "GMWI2 provides more direct relevance to subject phenotype following FMT treatment for IBS" but further evidence is required to conclusively demonstrate this statement across multiple independent datasets. Furthermore, to really demonstrate broad applicability of the GMWI2 index in this context it would be necessary to determining the ability to detect optimal FMT engraftment across broader disease states.

8. As suggested for the first FMT example, the authors should also include samples from multiple diet interventions and antibiotic treatments to demonstrate the broad applicability of the index in this context. These dietary intervention studies are obviously very different in bias and nature to intervention with large complex microbiome communities in FMT.

9. While the authors refer to a decrease in gut health associated with the reduced GMWI2 score when analysing stool samples it is important to remember the score is generated from a microbiome analysis of a stool provided from a healthy or diseased individual. There is no measure made of gut health or the actual microbial community within the gut. These compositional changes the model is trained upon could be the result of changes in transit time, stool consistency or many other factors not considered within the metadata. The manuscript should be revised to clearly highlight the association nature of the predictions and the conclusions worded to acknowledge these limitations.

Minor Comments:

In general the figures are difficult to follow, particularly Figure 1. The authors could consider presenting less text to improve interpretation and data quality.

1. In the introduction the authors state "incorporating data from all taxonomic ranks allows microbial signatures across various phylogenetic depths to be captured" - this should be reworded to recognise the critical distinction between taxonomy and phylogeny in this context

2. The authors should clarify the standardised bioinformatics pipelines does not remove all batch effects, only those associated with the bioinformatics analysis. Collection method, storage, extraction, etc. will still play a role.

3. The authors cite a publication [25] suggesting there are "on-going concerns about the long-term safety and efficacy of FMT in treating patients with chronic disease"; however, the evidence, including the conclusion from the cited study, suggests FMT is safe and efficacious for multiple diseases most notably UC a chronic condition. The authors should consider refocusing this statement.

Reviewer #1 (Remarks to the Author):

In this study, the authors introduced GMWI2 as a novel predictor of human health. GMWI2 was calculated as a weighted sum based on the presence/absence of microbial taxa identified from stool shotgun metagenomic data. The weights were determined through a LASSO-panelized logistic regression model using a pooled dataset comprising 8069 stool shotgun samples from existing literature. The predictive performance of the GMWI2 classification model was assessed on both the initial discovery dataset of 8069 samples and four additional external datasets.

Overall, this is a well designed study.

Author's Response: We thank the reviewer for this compliment!

One major technical concern is the potential risk of data leakage during model testing. The various validation procedures described in the section "Enhanced classification of healthy and non-healthy gut microbiomes with GMWI2" and the section "Evaluating the robustness of GMWI2 across different study populations" all employed a subset of the initial 8069 samples as test sets. However, the entire set of 8069 samples was used to compute the logistic regression coefficients and GMWI2 scores. Consequently, the GMWI2 scores inherently contain information from the test sets, irrespective of the chosen validation strategy. To mitigate the issue, the logistic regression coefficients and cutoff parameters should be estimated exclusively on the training data. While the authors did evaluate the GMWI2 model's performance on four external datasets, these evaluations did not specificially address the model's ability to distinguish between healthy and non-healthy samples.

Author's Response: Thank you for highlighting these important concerns. We appreciate the opportunity to clarify the following points:

The evaluation of GMWI2's capability to differentiate healthy (i.e., disease absence) from non-healthy (i.e., disease presence) gut microbiome samples can be conceptually divided into three phases. Only the second phase applies cross-validation strategies.

1. In the first phase, GMWI2 was trained and evaluated using the entire training set. More specifically, the entire set of 8069 samples was used to compute the logistic regression coefficients and GMWI2 scores. (The resulting coefficients of this model are displayed in **Fig. 2b**.) We performed this for the following reasons:
   - To identify the "theoretical limit" of our classifier's performance; in other words, what is the best it can perform?
   - To directly compare GMWI2 with other metrics (see **Fig. 3a**). Since one of our goals was to determine which of the five metrics performed best (even though we were testing on the training set), we did not feel compelled to implement cross-validation in this instance.

   The resulting GMWI2 scores of this model—which was tested back on the entire training set—are displayed in **Figs. 3a–c**. The resulting performances of this model in distinguishing between healthy and non-healthy samples are displayed in **Figs. 3a–e** and **4b**. In **Figs. 3e** and **4b**, the performances of this model are displayed under the labels "Training (GMWI2)" and "Training set", respectively.

2. In the second phase, GMWI2 was evaluated using cross-validation strategies. Importantly, these cross-validation strategies do not use all 8069 samples to train a classification model in each validation loop; and thus the models trained and evaluated in this second phase are completely different from those obtained in the first phase. Specifically, for each cross-validation strategy, a separate model is trained and evaluated for each train-test split, using scikit-learn's cross-validation functions. In each train-test split, the model is trained only on the "train" part of the split, and is evaluated only on the "test" part of the split. For each cross-validation strategy, the performances on the "test" parts of each split are then aggregated and summarized in the manuscript. Cross-validation (or Inter-study validation) performances are displayed in **Figs. 3e** and **4a–b**. In these figures, the cross-validation performances are displayed under the labels "LOOCV", "ISV", and "10-fold CV".

   ○ The cross-validation methods are demonstrated in the following subsection in the Jupyter Notebook in our code repository for reproducing the analyses in our manuscript, which can be run using Google Colab at this link: https://colab.research.google.com/github/danielchang2002/GMWI2/blob/main/manuscript/GMWI2_manuscript.ipynb#scrollTo=9b3631c3

3. In the third phase, the model obtained from the first phase is applied to four external (i.e., independent of the training set) datasets. The reviewer is correct in stating that "*these evaluations did not specificially address the model's ability to distinguish between healthy and non-healthy samples*", but rather, these evaluations demonstrate how GMWI2 (the one obtained by training on all 8069 samples, which is publicly available at https://github.com/danielchang2002/GMWI2) can be effectively applied to longitudinal datasets to track scores periodically following an intervention. The third phase is illustrated in **Fig. 5**.

Please see our added explanation throughout **page 13, lines 64–67** regarding these three phases of GMWI2 performance evaluation. To further ensure there is absolutely no ambiguity moving forward, we have provided a clear explanation of our methodology, as such:

> *"In line with standard protocols in cross-validation, the training of the GMWI2 model, including the computation of logistic regression coefficients, was confined strictly to the training partition of each train-test split of the total 8,069 samples."*

Lastly, we clarify each point raised by the reviewer:

1. The reviewer is indeed correct in noting that we "*employed a subset of the initial 8,069 samples as test sets*" and that "*consequently, the GMWI2 scores inherently contain information from the test sets, irrespective of the chosen validation strategy.*" As mentioned above in the

description of the first phase, the classifier's evaluation on the training set was intentionally done.

2. The more appropriate technique, where "*the logistic regression coefficients and cutoff parameters should be estimated exclusively on the training data,*" was subsequently implemented in **Figs. 3e** and **4a–b**.

3. There was no data leakage during any of our cross-validation (in **Figs. 3e** and **4b**) and inter-study validation (in **Figs. 4a–b**) procedures.

Another evidence of potential data leakage can be observed in the sentences discussing the selection of the regularization parameter C: "The selection of the regularization parameter C was achieved through hyperparameter tuning, where we evaluated various candidates and selected the value (C = 0.03) that yielded the optimum classification performance in inter-study validation (ISV)". The parameter C and other hyperparameters should be evaluated exclusively on the training set, along with logistic regression coefficients, for each given train-test split. The reestimation of these parameters helps ensure the robustness and integrity of the model's performance assessment.

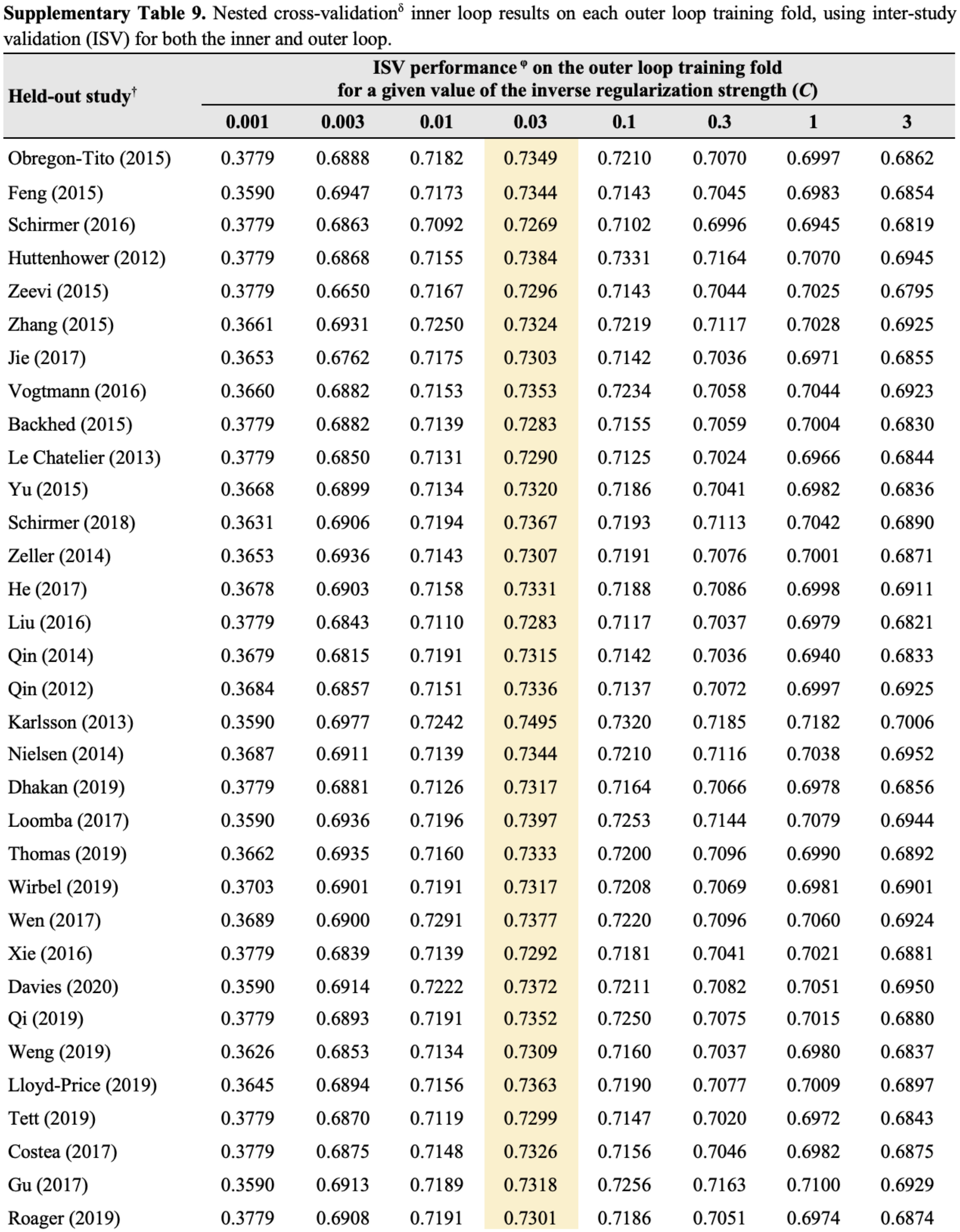
Author's Response: We are grateful to the reviewer for raising the crucial issue of potential data leakage during hyperparameter tuning. We concur with the principle that the regularization parameter *C*, along with other hyperparameters, must be calibrated solely within the training fold—in conjunction with logistic regression coefficients—for every individual train-test partition. (We are well aware that this practice is fundamental to preserving the robustness and integrity of the model performance evaluation.) Thus, following this principle, we clarify that we used *nested* cross-validation for hyperparameter selection in all cross-validation approaches demonstrated in our manuscript (i.e., 10-fold and leave-one-out in **Fig. 3e**; and ISV in **Fig. 4a**).

For hyperparameter selection for the final model trained on the entire dataset of 8,069 metagenomes, we further clarify that we used *nested* cross-validation that implements the inter-study validation (ISV) framework. (this specific point, which the reviewer pointed out, could have been explained more thoroughly in our original draft—thank you!) This method entails hyperparameter optimization confined exclusively to the training fold of each ISV loop (see modified **Supplementary Table 9** below). Our nested cross-validation protocol was as follows:

1. Outer loop: using ISV, establish the "leave-one-study-out" cross-validation structure to create multiple train-test splits.
2. Inner loop: for each outer loop train-test split, perform hyperparameter optimization within the training fold (or training set) by also using ISV.
3. For each outer loop train-test split, train the model using the optimized hyperparameters obtained from the inner loop on the training fold, and assess performance on the outer loop's corresponding test fold (or test set).

Consequently, we achieved an ISV accuracy estimation that is unmarred by hyperparameter data leakage.

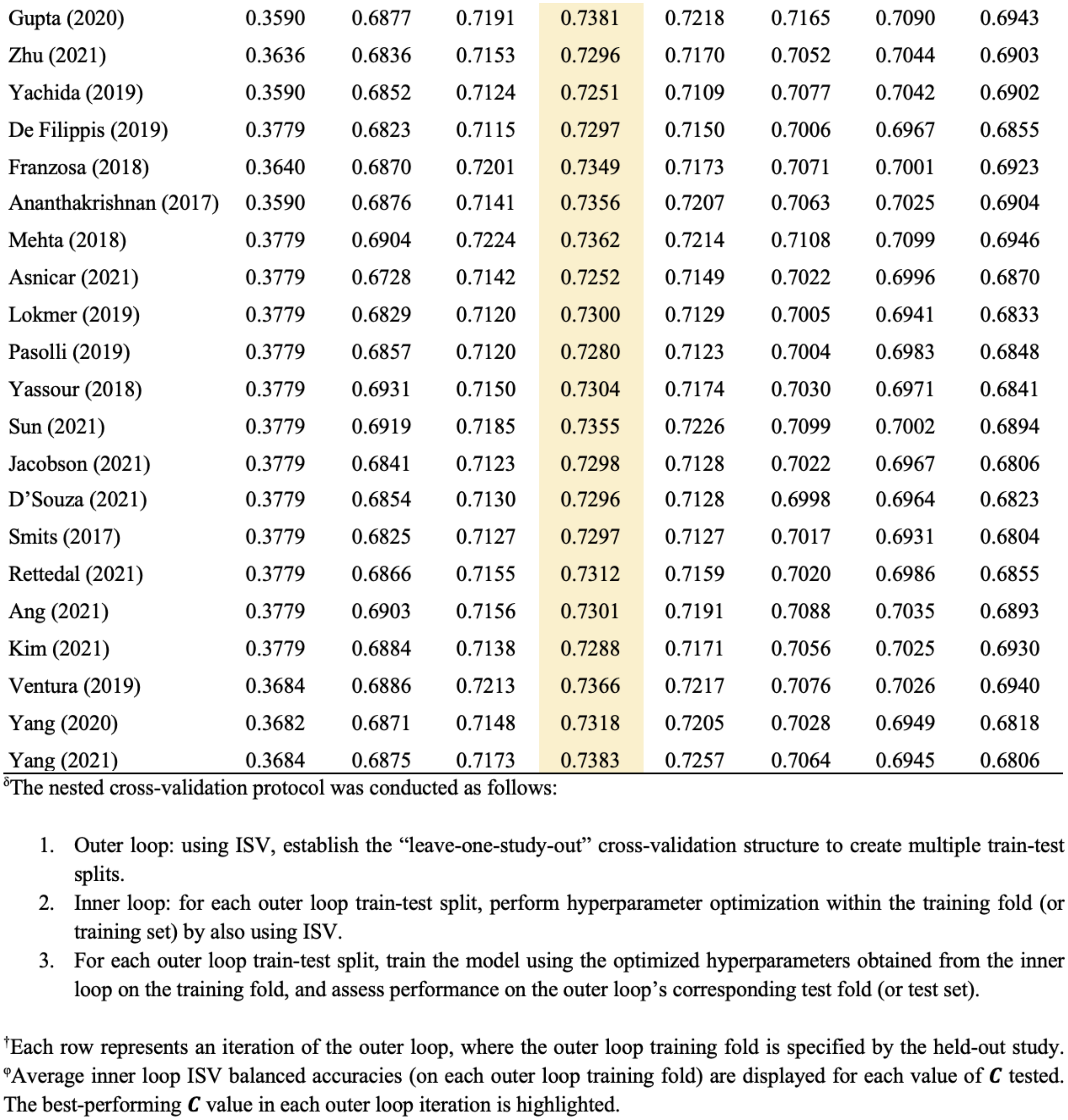
Our modified **Supplementary Table 9** (shown below) shows in detail the nested cross-validation inner loop results on each outer loop training fold, using ISV for both the inner and outer loop. Our results show that *C* = 0.03 consistently emerged as the optimal hyperparameter within each outer loop training fold, and hence, why we chose 0.03 as the inverse regularization strength (*C*).



Supplementary Table 9. Nested cross-validation[δ] inner loop results on each outer loop training fold, using inter-study validation (ISV) for both the inner and outer loop.

Held-out study[†]	ISV performance[φ] on the outer loop training fold for a given value of the inverse regularization strength (*C*)							
	0.001	**0.003**	**0.01**	**0.03**	**0.1**	**0.3**	**1**	**3**
Obregon-Tito (2015)	0.3779	0.6888	0.7182	0.7349	0.7210	0.7070	0.6997	0.6862
Feng (2015)	0.3590	0.6947	0.7173	0.7344	0.7143	0.7045	0.6983	0.6854
Schirmer (2016)	0.3779	0.6863	0.7092	0.7269	0.7102	0.6996	0.6945	0.6819
Huttenhower (2012)	0.3779	0.6868	0.7155	0.7384	0.7331	0.7164	0.7070	0.6945
Zeevi (2015)	0.3779	0.6650	0.7167	0.7296	0.7143	0.7044	0.7025	0.6795
Zhang (2015)	0.3661	0.6931	0.7250	0.7324	0.7219	0.7117	0.7028	0.6925
Jie (2017)	0.3653	0.6762	0.7175	0.7303	0.7142	0.7036	0.6971	0.6855
Vogtmann (2016)	0.3660	0.6882	0.7153	0.7353	0.7234	0.7058	0.7044	0.6923
Backhed (2015)	0.3779	0.6882	0.7139	0.7283	0.7155	0.7059	0.7004	0.6830
Le Chatelier (2013)	0.3779	0.6850	0.7131	0.7290	0.7125	0.7024	0.6966	0.6844
Yu (2015)	0.3668	0.6899	0.7134	0.7320	0.7186	0.7041	0.6982	0.6836
Schirmer (2018)	0.3631	0.6906	0.7194	0.7367	0.7193	0.7113	0.7042	0.6890
Zeller (2014)	0.3653	0.6936	0.7143	0.7307	0.7191	0.7076	0.7001	0.6871
He (2017)	0.3678	0.6903	0.7158	0.7331	0.7188	0.7086	0.6998	0.6911
Liu (2016)	0.3779	0.6843	0.7110	0.7283	0.7117	0.7037	0.6979	0.6821
Qin (2014)	0.3679	0.6815	0.7191	0.7315	0.7142	0.7036	0.6940	0.6833
Qin (2012)	0.3684	0.6857	0.7151	0.7336	0.7137	0.7072	0.6997	0.6925
Karlsson (2013)	0.3590	0.6977	0.7242	0.7495	0.7320	0.7185	0.7182	0.7006
Nielsen (2014)	0.3687	0.6911	0.7139	0.7344	0.7210	0.7116	0.7038	0.6952
Dhakan (2019)	0.3779	0.6881	0.7126	0.7317	0.7164	0.7066	0.6978	0.6856
Loomba (2017)	0.3590	0.6936	0.7196	0.7397	0.7253	0.7144	0.7079	0.6944
Thomas (2019)	0.3662	0.6935	0.7160	0.7333	0.7200	0.7096	0.6990	0.6892
Wirbel (2019)	0.3703	0.6901	0.7191	0.7317	0.7208	0.7069	0.6981	0.6901
Wen (2017)	0.3689	0.6900	0.7291	0.7377	0.7220	0.7096	0.7060	0.6924
Xie (2016)	0.3779	0.6839	0.7139	0.7292	0.7181	0.7041	0.7021	0.6881
Davies (2020)	0.3590	0.6914	0.7222	0.7372	0.7211	0.7082	0.7051	0.6950
Qi (2019)	0.3779	0.6893	0.7191	0.7352	0.7250	0.7075	0.7015	0.6880
Weng (2019)	0.3626	0.6853	0.7134	0.7309	0.7160	0.7037	0.6980	0.6837
Lloyd-Price (2019)	0.3645	0.6894	0.7156	0.7363	0.7190	0.7077	0.7009	0.6897
Tett (2019)	0.3779	0.6870	0.7119	0.7299	0.7147	0.7020	0.6972	0.6843
Costea (2017)	0.3779	0.6875	0.7148	0.7326	0.7156	0.7046	0.6982	0.6875
Gu (2017)	0.3590	0.6913	0.7189	0.7318	0.7256	0.7163	0.7100	0.6929
Roager (2019)	0.3779	0.6908	0.7191	0.7301	0.7186	0.7051	0.6974	0.6874

]



Gupta (2020)	0.3590	0.6877	0.7191	0.7381	0.7218	0.7165	0.7090	0.6943
Zhu (2021)	0.3636	0.6836	0.7153	0.7296	0.7170	0.7052	0.7044	0.6903
Yachida (2019)	0.3590	0.6852	0.7124	0.7251	0.7109	0.7077	0.7042	0.6902
De Filippis (2019)	0.3779	0.6823	0.7115	0.7297	0.7150	0.7006	0.6967	0.6855
Franzosa (2018)	0.3640	0.6870	0.7201	0.7349	0.7173	0.7071	0.7001	0.6923
Ananthakrishnan (2017)	0.3590	0.6876	0.7141	0.7356	0.7207	0.7063	0.7025	0.6904
Mehta (2018)	0.3779	0.6904	0.7224	0.7362	0.7214	0.7108	0.7099	0.6946
Asnicar (2021)	0.3779	0.6728	0.7142	0.7252	0.7149	0.7022	0.6996	0.6870
Lokmer (2019)	0.3779	0.6829	0.7120	0.7300	0.7129	0.7005	0.6941	0.6833
Pasolli (2019)	0.3779	0.6857	0.7120	0.7280	0.7123	0.7004	0.6983	0.6848
Yassour (2018)	0.3779	0.6931	0.7150	0.7304	0.7174	0.7030	0.6971	0.6841
Sun (2021)	0.3779	0.6919	0.7185	0.7355	0.7226	0.7099	0.7002	0.6894
Jacobson (2021)	0.3779	0.6841	0.7123	0.7298	0.7128	0.7022	0.6967	0.6806
D'Souza (2021)	0.3779	0.6854	0.7130	0.7296	0.7128	0.6998	0.6964	0.6823
Smits (2017)	0.3779	0.6825	0.7127	0.7297	0.7127	0.7017	0.6931	0.6804
Rettedal (2021)	0.3779	0.6866	0.7155	0.7312	0.7159	0.7020	0.6986	0.6855
Ang (2021)	0.3779	0.6903	0.7156	0.7301	0.7191	0.7088	0.7035	0.6893
Kim (2021)	0.3779	0.6884	0.7138	0.7288	0.7171	0.7056	0.7025	0.6930
Ventura (2019)	0.3684	0.6886	0.7213	0.7366	0.7217	0.7076	0.7026	0.6940
Yang (2020)	0.3682	0.6871	0.7148	0.7318	0.7205	0.7028	0.6949	0.6818
Yang (2021)	0.3684	0.6875	0.7173	0.7383	0.7257	0.7064	0.6945	0.6806

[δ]The nested cross-validation protocol was conducted as follows:

1. Outer loop: using ISV, establish the "leave-one-study-out" cross-validation structure to create multiple train-test splits.
2. Inner loop: for each outer loop train-test split, perform hyperparameter optimization within the training fold (or training set) by also using ISV.
3. For each outer loop train-test split, train the model using the optimized hyperparameters obtained from the inner loop on the training fold, and assess performance on the outer loop's corresponding test fold (or test set).

[†]Each row represents an iteration of the outer loop, where the outer loop training fold is specified by the held-out study.
[φ]Average inner loop ISV balanced accuracies (on each outer loop training fold) are displayed for each value of $\boldsymbol{C}$ tested. The best-performing $\boldsymbol{C}$ value in each outer loop iteration is highlighted.

In all, nested cross-validation in our ISV procedure is designed to preserve the integrity of our model's performance assessment during hyperparameter selection for the final model trained on the entire dataset of 8,069 metagenomes.

We revised the original text accordingly on **page 35–36, lines 480–489** of our revised manuscript:

> **Before:** *"A Lasso-penalized logistic regression model (Python library "scikit-learn" v1.0.2) was trained on the binary presence/absence taxonomic profiles of the entire pooled dataset to predict disease presence. The L1 (Lasso) penalty was utilized with the LIBLINEAR solver[50]. The random state was set to 42, the regularization parameter C to 0.03, and class weight to "balanced". The selection of the regularization parameter C was achieved through hyperparameter tuning, where we evaluated various candidates and selected the value (C =*

> *0.03) that yielded the optimum classification performance in inter-study validation (ISV) (**Supplementary Table 9**). The class weight was set to "balanced" in order to account for the unbalanced class proportions in our pooled dataset."*
>
> **After:** "*A Lasso-penalized logistic regression model (Python library "scikit-learn" v1.0.2) was trained on the binary presence/absence taxonomic profiles of the entire pooled dataset of 8,069 metagenomes to predict disease presence. The L1 (Lasso) penalty was utilized with the LIBLINEAR solver[50]. The random state was set to 42, and the class weight was set to "balanced" in order to account for the unbalanced class proportions in our pooled dataset. Hyperparameter tuning—specifically the selection of the regularization parameter C—was achieved through nested cross-validation that implements the inter-study validation (ISV) framework. Herein, we evaluated various candidates and selected the value that yielded the optimum classification performance in ISV (**Supplementary Table 9**; see table footnote for our nested cross-validation protocol). C = 0.03 consistently emerged as the optimal hyperparameter within each outer loop training fold and was thus selected for the final GMWI2 model."*

In addition, the authors used balanced accuracy (averaged proportion of correctly classified healthy and non-healthy samples) to evaluate the performance of the GMWI2 classification model. However, the model can output a third category, "not determined"(or defer), in addition to classifying samples as healthy or non-healthy. To gain a comprehensive understanding of the model's performance, it would be essential to know, for each validation, the number or percentage of samples classified as "not determined". In extreme cases, it is possible for the model to achieve 100% balanced accuracy while having limited practical value. For example, the model may correctly predict 1 healthy sample and 1 unhealthy sample but unable to determine the label for the remaining 98 samples.

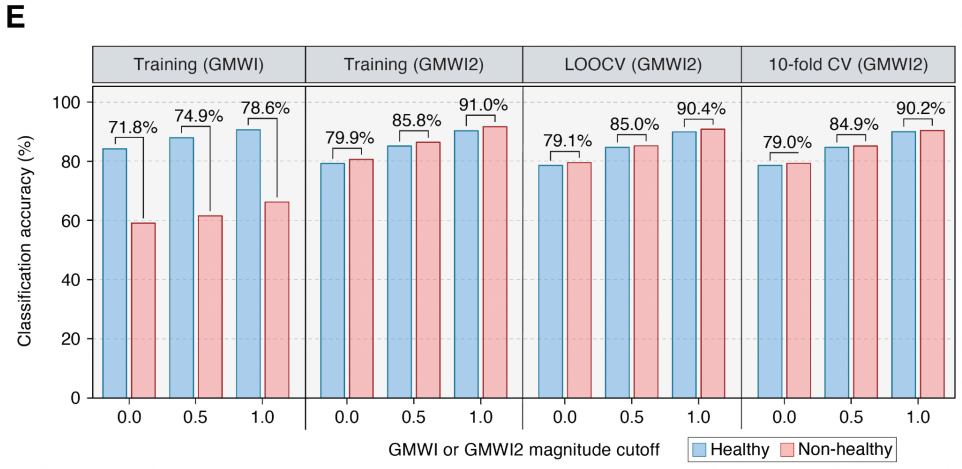
Authors' Response: We appreciate the opportunity to address each of the points raised:

On the point of "*However, the model can output a third category, "not determined" (or defer), in addition to classifying samples as healthy or non-healthy,*" we clarify that our current methodology does not inherently categorize gut microbiome samples into a third option. GMWI2 yields a continuous score, where the sign (negative or positive) is indicative of disease presence or absence, respectively; and higher magnitudes imply greater confidence in the prediction. The "not determined" (or "defer") category is an optional feature, applicable when a user decides to implement a GMWI2 magnitude cutoff. Scores falling below this user-defined cutoff (e.g., between –1.0 and +1.0) can be classified as "defer." This aspect is discussed in detail on **page 37, lines 515–520**:

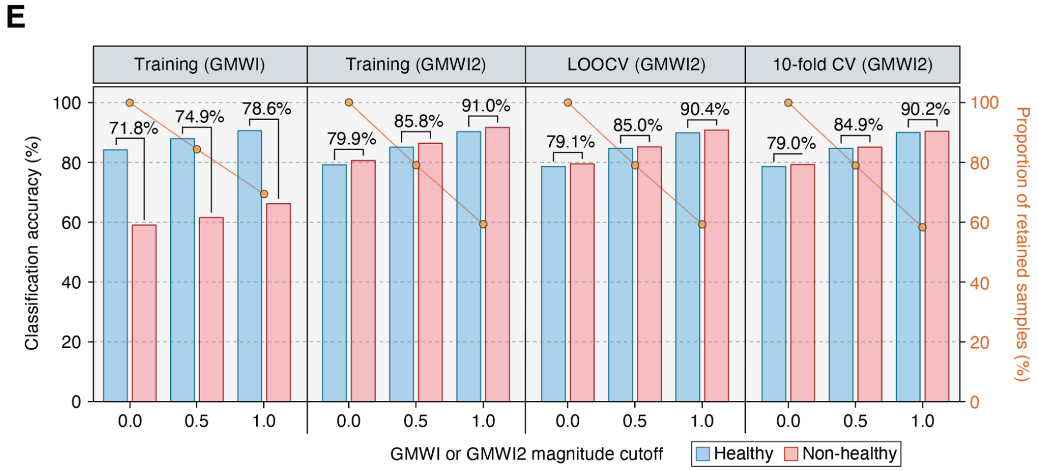
> *"Of note, our current methodology does not inherently categorize gut microbiome samples into a third option. GMWI2 yields a continuous score, where the sign (negative or positive) is indicative of disease presence or absence, respectively; and higher magnitudes imply greater confidence in the prediction. The "defer" (or "not determined") category is an optional feature, applicable when a user decides to implement a non-zero GMWI2 magnitude cutoff c. Scores falling below this user-defined cutoff (e.g., between –1.0 and +1.0) can be classified as "defer.""*

Regarding "*To gain a comprehensive understanding of the model's performance, it would be essential to know, for each validation, the number or percentage of samples classified as 'not determined',*" we have already conducted an analogous analysis with the same aim, illustrated in **Fig. 3d**, for all training data samples. While this analysis was not specific to each training-test split, it effectively demonstrates the inverse relationship between classification accuracy and the number of samples eligible for class prediction. To further clarify this point, we have included additional data in **Fig. 3e**, now showcasing the percentage of samples retained at GMWI2 cutoffs of 0 (no cutoff), 0.5, and 1.0 (as the reviewer has requested). The modified **Fig. 3e** is presented as follows:

Before:



After:



In addition, we have added the following sentence to the figure legend: *"Orange points represent the proportion of retained samples (y-axis, right) for the corresponding index magnitude cutoff."*

In response to the concern that "*In extreme cases, it is possible for the model to achieve 100% balanced accuracy while having limited practical value. For example, the model may correctly predict 1 healthy sample and 1 unhealthy sample but unable to determine the label for the remaining 98 samples,*" we acknowledge this scenario. However, we ensure that such a situation would occur only under extraordinarily extreme circumstances and if the user deliberately chooses to set it up in this manner.

Here, it is very important to note that our model allows users to select their desired GMWI2 magnitude cutoff based on their confidence level preference in the predictions. This user-driven approach, which offers flexibility between high confidence in a limited dataset and broader range predictions with lesser confidence, is a distinctive advantage of our method over traditional binary classification machine learning techniques. This is further elaborated on **pages 29–30, lines 345–356**:

> *“In our analyses in which we incrementally increased the GMWI2 magnitude cutoff, we recognize an inverse relationship between classification accuracy and the volume of samples eligible for class prediction. Therefore, constraining this magnitude cutoff to a single, optimal value may not be universally applicable; instead, the selection of this parameter should be flexible and determined by the user, tailored to the specific context and acceptable accuracy thresholds of their individual datasets. In other words, users can select their desired GMWI2 magnitude cutoff based on their confidence level preference in the predictions. This user-driven approach, which offers flexibility between high confidence in a limited dataset and broader range predictions with lesser confidence, is a distinct advantage of our method over traditional binary-output machine learning techniques. Moreover, our findings thus foster the potential utility of a “reject option”[41,42] for low GMWI2 magnitudes, which can serve as a criterion to redirect relatively uncertain predictions to other screening methods—this concept captures the understanding that certain aspects of health and disease are not fully explainable solely by the gut microbiome.”*

I also have some personal comments on the significance of the study. The authors claimed that the model "pave the way for the early screening of adverse gut health shifts" and is "adaptable for researchers interested in the translational applications of human gut microbiome science". While these claims hold considerable promise, I encountered some difficulty in understanding the practical utility of the model in translational applications. If an individual uses this tool and the output indicates an "unhealthy" status, the interpretation becomes a pivotal question. Does an "unhealthy" result indicate that the individual is currently experiencing health issues or that the individual may develop one of the 12 diseases you studied in the future? Furthermore, what actions should the individual take in response? From a practical standpoint, if the person is already displaying symptoms, the microbiome test may not provide additional value in confirming their health status. On the other hand, if the individual is asymptomatic, he/she may have to wait until symptoms manifest before taking any proactive measures. This implies that merely categorizing individuals as healthy or unhealthy may not offer actionable guidance.

Authors’ Response: We are grateful to the reviewer for their comments, which prompted a thorough reevaluation of how we present our index’s translational utility. This reflection has been instrumental in enhancing our manuscript’s clarity and applicability in the field. We now address all points mentioned above:

1. "*While these claims hold considerable promise, I encountered some difficulty in understanding the practical utility of the model in translational applications*."

   We respectfully request that the reviewer carefully review the final paragraph of our Discussion section. Here, we delve into the innovative potential of our relatively novel approach, with an anticipation of its further development and broader application in the future.

   > *"To conclude, GMWI2 is not for confirming a specific diagnosis of a disease, but rather to function as the proverbial "canary in a coal mine" by serving as an early warning system. It is designed to detect potentially adverse shifts in overall gut health, which could inform dietary or lifestyle modifications to prevent mild issues from escalating into severe health conditions, or prompt further diagnostic tests for a more detailed and accurate diagnosis. This could be particularly useful in clinical scenarios such as selecting FMT/organ donors, where gut health could be indicative of overall health, or in assessing early predictors of disease flare-ups. In conditions like rheumatoid arthritis and other autoimmune inflammatory disorders, GMWI2 could guide decisions on which patients might benefit from tapering or stopping therapy. In this sense, GMWI2 may potentially usher in a novel, transformative era in gut microbiome-centric health analytics, allowing for nuanced health evaluations tailored to individual microbial signatures. Looking ahead, integrating GMWI2 into a larger decision network alongside other biomeasurements (e.g., multi-omics, wearables) and AI-based models has the potential to open up exciting possibilities for healthy aging[43] and preventative healthcare and wellness strategies[44,45], driven by insights from our gut microbiome."*

   In addition, we would like to mention that the understanding of what constitutes a deviation from "normal" health, and learning how to rapidly and robustly detect such deviations, are highly important areas of study in academic medicine today. Without a doubt, precisely addressing these challenges will help accelerate the development of new technologies for detecting early signs of disease *prior to the occurrence of specific, diagnosable (visible) symptoms.* Therefore, the creation of algorithm-driven markers that can infer one's general health state, especially from biospecimens that can be collected regularly and non-invasively, is a very promising avenue forward. And with further development, our gut microbiome-based predictor could serve as an appealing contribution toward comprehensive medical and preventive health screening programs.

   Results from such tests can then serve as an entry point for follow-up tests and procedures. In this sense, "generalized alarm bells" goes beyond merely having "some theoretical utility", and are just as in need as novel diagnostics designed for a particular disease. Both goals are not mutually exclusive and should be pursued together. And yes, identifying the specific malady is crucial once symptoms arise, but that is not what is being sought after in this study (that would be the next step). In our view, we demonstrate unequivocally the proof-of-concept that our health index—based on a snapshot of the gut microbiome—could have practical merit in the clinical setting, and we fully stand by the results and conclusions of this study.

Lastly, let's consider an example regarding financial credit scores in the United States. Say, after a long steady period of having a good credit score (750+), one suddenly receives a report of 450. Assuming this person is a financially responsible being, this precipitous drop in score would certainly lead one to check her/his latest bank statements, credit card bills, mortgage payments, etc. Hence, although the cause for a significant drop in credit score is initially unclear, it would nonetheless spur action to find out the why and what to do next. Analogously in the case of maintaining wellness and preventing disease, our hope is that GMHI may one day serve as a tool for turning microbiome data into actionable information.

We hope these descriptions effectively convey the *"practical utility of the model in translational applications*" that the reviewer seeks.

2. "*If an individual uses this tool and the output indicates an "unhealthy" status, the interpretation becomes a pivotal question. Does an "unhealthy" result indicate that the individual is currently experiencing health issues or that the individual may develop one of the 12 diseases you studied in the future?*"

   GMWI2 is a test for overall gut health, and is not meant to diagnose disease (see explanation directly above). An "unhealthy" result does not necessarily mean that one has visible health issues nor that the individual will develop one of the twelve diseases. Instead, this result is a summary statistic describing shifts in taxonomic composition towards known non-healthy gut microbiome states. It is not the case that our model can only predict if an individual has one of the twelve diseases, but rather our model can generalize to predicting overall gut health or dysbiosis. (We acknowledge that the term "dysbiosis" is poorly defined in the literature, but we nonetheless attempt to provide a reasonable, quantitative measure based on a massive corpus of real gut microbiome data.)

3. *"Furthermore, what actions should the individual take in response?"*

   As we mentioned in our Discussion section, GMWI2 is designed to detect potentially adverse shifts in overall gut health, which could inform dietary or lifestyle modifications to prevent mild issues from escalating into severe health conditions, or prompt further diagnostic tests for a more detailed and accurate diagnosis. Suppose an individual is planning on implementing various long-term lifestyle changes (or, in a different case, use a long-term therapeutic drug) to improve health and well-being. With further development past this prototype stage, GMWI2 can be envisioned to serve as a barometer for quantitative real-time monitoring of the health effects of various lifestyle modifications. Who would not want this?!

4. *"From a practical standpoint, if the person is already displaying symptoms, the microbiome test may not provide additional value in confirming their health status."*

   Of course, if someone is already displaying symptoms (e.g., enteric inflammatory flares related to Crohn's Disease), then they do not need to consult their GMWI2 score to know that they need to acutely treat their symptoms!

5. *“On the other hand, if the individual is asymptomatic, he/she may have to wait until symptoms manifest before taking any proactive measures.”*

   This addresses our point and highlights what needs to be improved in today’s practice of medicine—the approach of waiting for symptoms to appear before taking action. Wouldn't it be beneficial to provide a way to take proactive measures before symptoms manifest, especially if an individual is asymptomatic? Signs of decreased gut health can serve as an indicator (of possibly several) to perform additional tests to detect silent yet insidious conditions.

6. *“This implies that merely categorizing individuals as healthy or unhealthy may not offer actionable guidance.”*

   Our objective is not to offer highly specific, individually tailored actionable guidance or a “one-size-fits-all” solution, as these are best left to licensed medical professionals. Instead, our tool is designed to offer a novel method for dynamically monitoring an individual’s health in a semi-real-time manner through the analysis of associative gut microbiome signatures. As discussed in our manuscript and our previous responses, this capability to assess health can be immensely valuable for individuals focused on optimizing their well-being through practical dietary or lifestyle modifications—spurred by data on their microbiome!

Finally, I want to express my reservation regarding the following statement "Any additional information required to reanalyze the data reported in this paper is available from the corresponding author upon reasonable request." In my opinion, this statement may not be acceptable. At the minimum, the microbial taxonomic profiles of the 8069 samples should be made available for download. Without access to these essential data, the reproducibility of the results presented in this manuscript cannot be rigorously accessed.

Authors’ Response: We would like to clarify. *“the microbial taxonomic profiles of the 8069 samples”* that the reviewer requests were already provided in our GitHub repository prior to our original submission. The wording of *“Any additional information required to reanalyze the data reported in this paper is available from the corresponding author upon reasonable request.”* was in regard to anything else the future reader may want to know besides all the essential data already provided in either our Supplementary Tables (or Data) or our GitHub repository. But many thanks for checking!

We have provided the following on **page 38, lines 539–542** to make sure that our stance on “open science” is clear:

> *“The source code for the tool, processed datasets (including the microbial taxonomic profiles of all metagenome samples analyzed in this study), and code notebooks essential to reproduce all results presented in our study, as well as complete instructions for installation and usage, are freely available online at https://github.com/danielchang2002/GMWI2.”*

Minor comments:

1. Line number is missing. Page and line numbers would make it a lot easier for me to provide feedback."

Authors' Response: Thank you for pointing this out. These are now included.

2. "...long-term safety and efficacy of FMT in treating patients with chronic diseases, computational tools like GMWI2 could be useful in assisting with the selection of healthy donors and stool samples." How does GMWI2 address the safety issue? The taxa with negative coefficients do not include many invasive pathogens such as Enterococcus faecium/faecalis.

Authors' Response: Thank you for pointing that out. To prevent any confusion, we have omitted the previously mentioned line from our manuscript.

We wish to clarify that being of use in the selection of healthy donors and stool samples for FMT does not necessarily equate to resolving all safety concerns associated with FMT. It's important to recognize that a therapy deemed "safe" in terms of causing no harm may not necessarily lead to clinical improvement in patients. (This is a common observation in post-Phase I clinical trials.) Our intention was to highlight that computational tools like GMWI2 can be instrumental in addressing the intricate challenges of selecting donor stools for FMT. We have clarified this point on **page 24, lines 259–262**:

> *"In light of the clinical significance and the complexities involved in donor screening for FMT[24,25], computational tools such as GMWI2 could help in guiding the selection of suitable healthy donors and their stool samples."*

Additionally, as detailed in **Supplementary Data 4**, our analysis reveals that taxa with negative coefficients in our model do include some opportunistic pathogenic taxa, like various *Clostridium* species. While recognized pathogens such as *E. faecium/faecalis* may not exhibit negative coefficients in our model, it is important to note that pathogenic traits are much more accurately identified at the strain level, which is beyond our model's scope. Moreover, it is well-documented that not all disease-related gut microbiomes contain invasive pathogens. We mention this on **page 30, lines 372–379**:

> *"Fifth, our analysis revealed that well-known pathogens, including Enterococcus faecium/faecalis, did not display negative coefficients in our GMWI2 framework. Nevertheless, we did observe negative coefficients for certain opportunistic pathogenic taxa, notably among various Clostridium species, as detailed in **Supplementary Data 4**. It is important to emphasize that the determination of pathogenic traits is more accurately conducted at the strain level, which falls outside the scope of our model. Additionally, it is widely acknowledged that not every gut microbiome associated with chronic, non-communicable disease necessarily harbors invasive pathogens."*

3. It is unclear how the cutoff parameter c was determined for each validation strategy. Is there a single, universal c derived from the 8069 samples and applied uniformly across all scenarios, or is this parameter reestimated for each train-test split? In addition, the values of c need to be reported.

Authors' Response: Thank you for your question regarding a single, universal c for the GMWI2 score (in magnitude) cutoff parameter. We'd like to clarify that there is no "derived" c from the training data of 8069 samples, or from each train-test split; but rather something much simpler. The three values we had tried in **Figs. 3d–e**, i.e., 0, 0.5, and 1, are merely example values to demonstrate the concept of the "GMWI2 magnitude cutoff"; and also to illustrate that the inherent trade-off between higher accuracy (by increasing the magnitude cutoff) and the consequent exclusion of potentially valuable samples is a key consideration in our methodology. The three values (0, 0.5, and 1.0) are just suggestions for interpreting our logistic regression classification results.

To make sure readers are not confused, we did a better job in introducing the motivation of this analysis on **page 16, lines 108–114** as follows:

> *"The results presented in **Fig. 3c** of our study revealed a notable trend. Specifically, when GMWI2 (and GMWI) scores exhibit a more positive or negative value, there is a corresponding increase in the proportion of actual healthy and non-healthy samples, respectively. This trend suggests a potential increase in the confidence of phenotype classification. In contrast, as these values near zero, our confidence in accurately determining the presence or absence of a disease decreases. To examine this point more closely, we next investigated how setting a minimum GMWI2 threshold or cutoff parameter could enhance classification accuracy for phenotype prediction."*

We also mention a brief clarification on **page 16, lines 119–120**:

> *"(these cutoffs are examples to illustrate the concept of the GMWI2 magnitude cutoff.)"*

Reviewer #2 (Remarks to the Author):

Understanding the gut microbiome and its role in disease represents an ongoing challenge for the research field. The ability to rapidly and efficiently assess a participants microbiome as health associated or diseased has the potential to provide key insights and understandings in many contexts. Too often statements such as dysbiotic are used without clear definition or quantification limiting comparison between individuals, cohorts and studies more broadly. The authors present an refined version of their previously published index, the Gut Microbiome Wellness Index (GMWI2) to address this challenge.

The presented work represents an advance in standardising reporting of microbiome composition in stool samples and linking this to general patient state; however, there are some key limitations both in terms of the overall approach, the specific implementation and the limits in comparison. Most notably the comparisons being made to only existing GMWI and traditional diversity measures rather than current best practice analysis methods. Very few high quality studies, including many of those used as the exemplars rely purely on diversity to determine biological relevance of the findings. These additional comparisons should be included and this situation more clearly articulated in the manuscript to ensure appropriate application of the developed index. Further specific considerations are included below:

Authors' Response: We seek clarification and/or specific examples of the *"best practice analysis methods"* and the specific meaning of *"these additional comparisons"* mentioned by the reviewer. We kindly request details on the alternative methods for defining gut health, apart from the GMWI and alpha-diversity metrics already presented in our manuscript (see **Fig. 3A** regarding our data using three different types of alpha-diversities). If the reviewer's proposed methods are relevant within the scope of our study and provide a meaningful contribution, then we are certainly open to incorporating them into our analysis.

Furthermore, a key question concerning alpha-diversity metrics is whether they sufficiently provide insights into the general health state. If so, what precisely defines "high" alpha-diversity, and what is considered "low"? Given the scarcity of answers in the existing literature, our work aims to address this gap. Moreover, our main goal extends beyond merely examining the biological relevance of findings.

1. The choice of samples for inclusion and testing is of critical importance to the model overall. How does the definition of healthy impact the overall accuracy and robustness of the model and which factors defining health (e.g. BMI, age, etc.) contribute most strongly to the prediction.

Authors' Response: We appreciate this question. The selection of samples for inclusion and testing is indeed critical. To minimize potential biases, we earnestly sought to identify as many samples as possible and then selected those that met our sample/study exclusion criteria, which were consistently applied across all studies and samples.

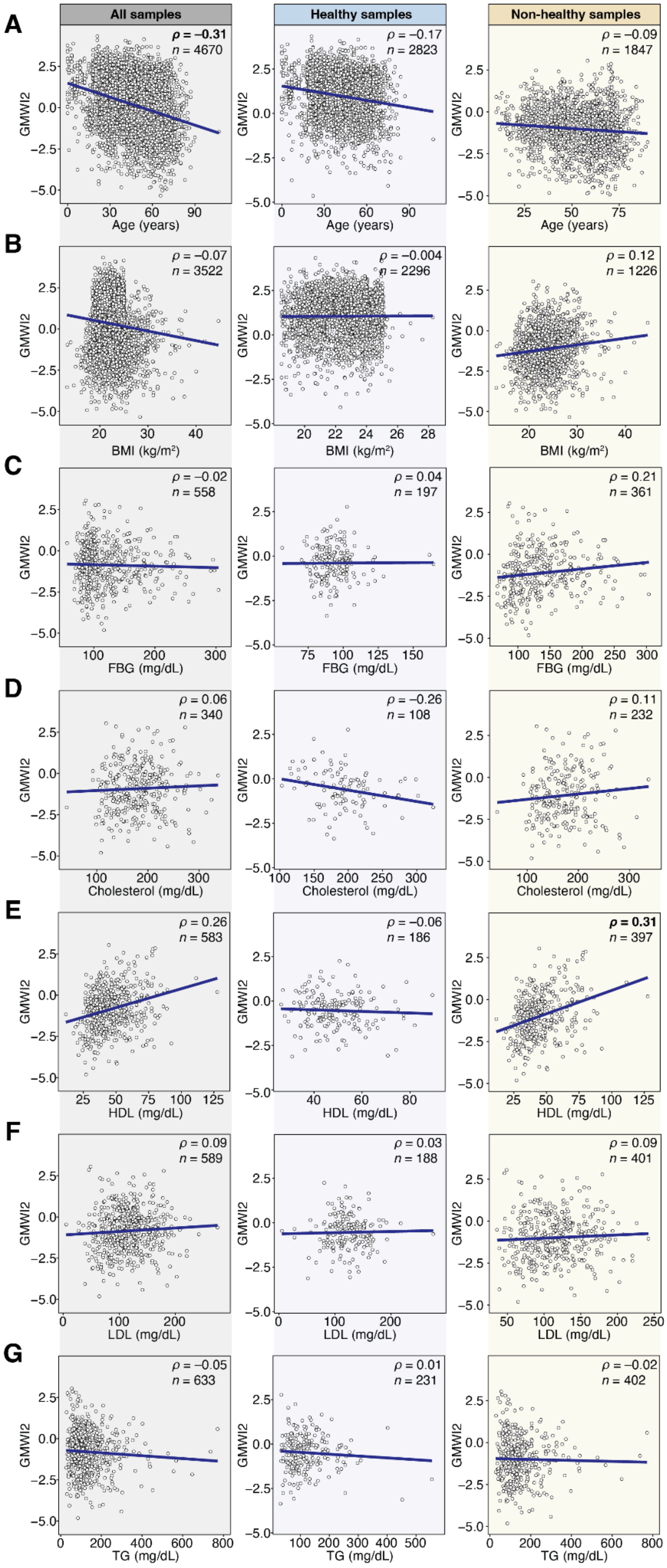
Importantly, we maintained the same definition of “healthy” as in our previous study (i.e., self-reported absence of any disease or disease-related symptoms) where we introduced the first version of GMWI (DOI: 10.1038/s41467-020-18476-8). This consistency allows for a fair comparison of performance between the two versions of our index. Moreover, since there is no universally accepted definition of “healthy”, it is prudent to adhere to a reasonable standard consistently.

Regarding which factors defining health (e.g., BMI, age) contribute most significantly to the prediction, we analyzed which of these are most strongly associated with GMWI2. The rationale here is that clinical or demographic factors strongly correlated with GMWI2 likely play a critical role in gut microbiome-based classification outcomes. In the extreme case, a perfect correlation with GMWI2 would indicate that it is as informative in distinguishing healthy vs. non-healthy individuals as the gut microbiome itself.

Our findings are presented in the new **Supplementary Figure 2** below. (*ρ* represents Spearman’s correlation coefficient; *n* denotes the number of study participants with available clinical or demographic information; FBG stands for fasting blood glucose; HDL for high-density lipoprotein; LDL for low-density lipoprotein; and TG for triglycerides.) As depicted in these correlation scatterplots between GMWI2 and various clinical/demographic characteristics, most variables exhibit weak correlations with GMWI2. Therefore, we find little evidence that any single clinical or demographic factor alone can reliably separate healthy from non-healthy individuals.

We have also mentioned these findings on **page 15, lines 93–96**:

> *“Lastly, we observed weak correlations between GMWI2 and clinical/demographic characteristics (|Spearman’s ρ | < 0.3;* ***Supplementary Fig. 2****), indicating that these factors do not significantly influence gut microbiome-based classification outcomes.”*



Supplementary Figure 2. Correlations between GMWI2 and clinical/demographic characteristics. All variables show weak correlations with GMWI2. ρ, Spearman's correlation coefficient; n = number of study participants with available clinical or demographic information; FBG, fasting blood glucose; HDL, high-density lipoprotein; LDL, low-density lipoprotein; and TG, triglycerides.

2. While small size studies were reasonably excluded from the training set, however, is there an observable difference when evaluating robustness of the model of these samples. If this model is to be broadly applied it is essential to demonstrate it is equally applicable on small datasets more typical of those it is likely to be applied to. Further assessments of these types of datasets would be highly valuable in demonstrating usefulness of the approach.

Authors' Response: Thank you for raising this question. In our comprehensive review of human stool metagenome studies, we identified that several studies comprised fewer than 40 samples, which was one of our study exclusion criteria. We made a methodological decision to exclude these studies from our analysis, considering the potential limitations in the robustness and reliability of "pilot-scale" microbiome studies, especially in the context of current-day standards. We clarify this on **page 33, lines 419–421**:

> *"Studies with fewer than 40 samples were also excluded from our analysis, considering the potential limitations in robustness and reliability of microbiome data from such "pilot-scale" microbiome studies."*

The reviewer's comments on the broad applicability of GMWI2 are highly relevant. This is the primary reason we implemented inter-study validation (ISV). ISV is crucial as it demonstrates the significant variability in classification performance that can arise depending on the chosen validation set. In essence, it provides a range of classification accuracies achievable when applying GMWI2 across 54 independent validation sets. Notably, ISV is designed to be indifferent to the size of the held-out study. **Fig. 4a** specifically showcases the performance of GMWI2 on various held-out studies, along with details on their respective sample sizes. We clarify this and acknowledge the reviewer's points on **page 20, lines 166–177**:

> *"Although studies with small sample sizes were excluded from the training set (see study exclusion criteria in **Fig. 1a** and **'Methods'**), in general, it is crucial to validate any classification model across datasets of varying sample sizes[19]. Conducting further assessments on gut microbiome datasets of different scales would demonstrate the broad applicability of GMWI2. To this end, we conducted inter-study validation (ISV) to assess the impact of batch effects (i.e., technical or biological variations associated with study populations) on GMWI2 performance stability. In this approach, we iteratively excluded a single study, trained the GMWI2 model on the remaining studies, and evaluated its classification performance on the excluded study[22]. (The excluded study essentially becomes the independent cohort.) ISV is crucial as it demonstrates the significant variability in classification performance that can arise depending on the choice of validation set. In essence, it provides a range of classification accuracies achievable when applying GMWI2 across 54 independent validation sets. Notably, ISV is designed to be indifferent to the size of the held-out studies."*

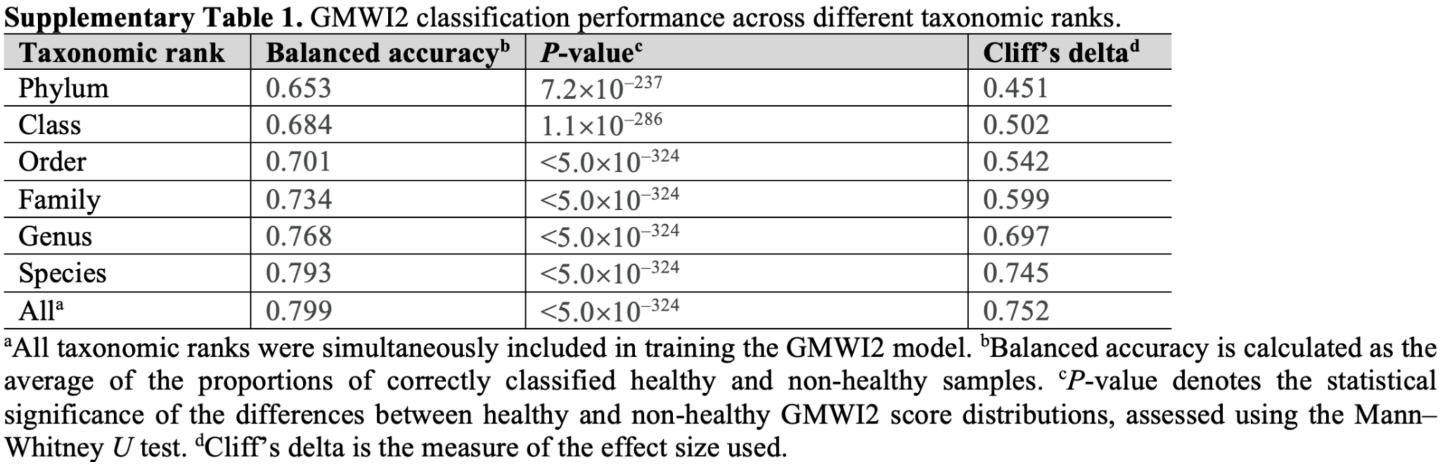
Fig. 4a specifically showcases the performance of GMWI2 across the full range of held-out studies, along with details on their sample sizes. Despite the notable variation in classification performance across different studies (see gold points indicating ISV classification accuracy per study in **Fig. 4a** and **Supplementary Table 3**), the average balanced accuracy was 75.8%. This performance rose to

86.9% when considering samples with GMWI2 scores lower than –1 or higher than 1. In all, our analysis revealed no discernible correlation between the model's predictive performance and the sample size of the held-out datasets.

3. It is notable that multiple taxonomic levels were identified with non-zero coefficients. Given the dependence of the taxonomic hierarchy what impact does this have on the model?

Authors' Response: Thank you for raising this important point, which we had, coincidentally, contemplated previously. The interdependence within and across different levels of the taxonomic hierarchy does indeed introduce multicollinearity, which can pose challenges in the interpretation of regression coefficients. However, it's worth noting that our comprehensive approach, which encompasses all taxonomic levels, had yielded better performance compared to the inclusion of a single taxonomic level, such as species (see new **Supplementary Table 1** below). Given our primary focus on optimizing classification performance, we have chosen to prioritize this aspect, which has led us to set aside the multicollinearity concern in this context. We have mentioned this valuable point on **pages 12–13, lines 48–54**:

> *"It is worth mentioning that several taxonomic levels exhibited non-zero coefficients in our analysis. This is likely due in part to the interdependence across different levels of taxonomic hierarchy introducing multicollinearity, which complicates the interpretation of regression coefficients. However, our approach in encompassing all taxonomic levels demonstrated higher classification performance compared to when using only a single taxonomic level (**Supplementary Table 1**). Given that our primary objective of optimizing classification accuracy, we chose to prioritize this aspect, leading us to set aside the multicollinearity concern."*



Supplementary Table 1. GMWI2 classification performance across different taxonomic ranks.

Taxonomic rank	Balanced accuracy[b]	*P*-value[c]	Cliff's delta[d]
Phylum	0.653	7.2×10^{-237}	0.451
Class	0.684	1.1×10^{-286}	0.502
Order	0.701	$<5.0\times10^{-324}$	0.542
Family	0.734	$<5.0\times10^{-324}$	0.599
Genus	0.768	$<5.0\times10^{-324}$	0.697
Species	0.793	$<5.0\times10^{-324}$	0.745
All[a]	0.799	$<5.0\times10^{-324}$	0.752

[a]All taxonomic ranks were simultaneously included in training the GMWI2 model. [b]Balanced accuracy is calculated as the average of the proportions of correctly classified healthy and non-healthy samples. [c]*P*-value denotes the statistical significance of the differences between healthy and non-healthy GMWI2 score distributions, assessed using the Mann–Whitney *U* test. [d]Cliff's delta is the measure of the effect size used.

4. While the data is presented comparing the GMWI2 to more basic measures (e.g. diversity) using the training dataset, further comparisons on independent sample sets to more accurately define where the application of the GMWI2 model provides maximum value would be highly advantageous.

Authors' Response: Thank you for your suggestion to demonstrate the predictive capability of GMWI2 on independent sample sets. We already recognized the importance of robust validation, and therefore why we implemented the interstudy validation (ISV) analysis in **Figure 4**, as noted above. (this analysis was specifically designed to assess the performance of GMWI2 across different studies and datasets.)

Additionally, in response to your comment and to further reinforce the validity of our model, we have compiled an independent validation cohort consisting of 1,140 samples derived from six published studies. This cohort includes stool metagenomes from healthy individuals as well as patients diagnosed with ankylosing spondylitis, pancreatic cancer, or Parkinson's disease. All metagenome samples in this dataset were classified into one of two categories: healthy or non-healthy. We believe this approach addresses your insightful feedback and substantiates the robustness and utility of the GMWI2 model.

GMWI2 scores on the external validation set are shown in the boxplots in our new **Figure 5**:

Figure 5

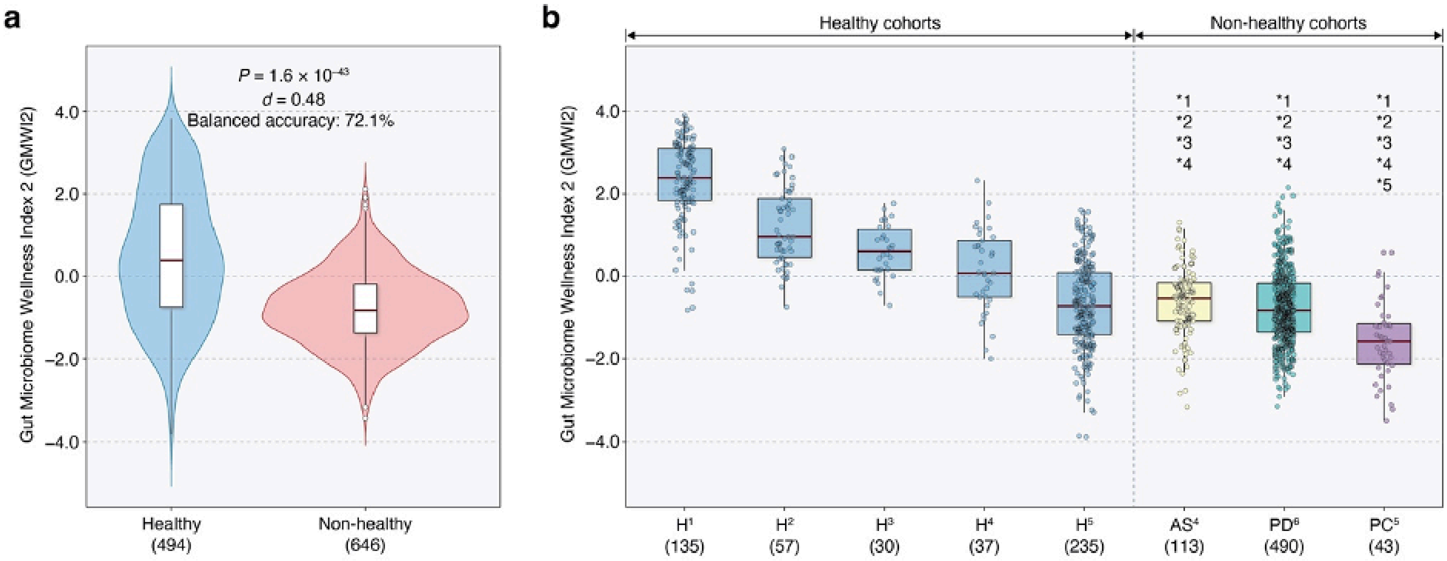


Figure 5. Comparative analysis of GMWI2 in healthy and non-healthy independent validation cohorts. (a) GMWI2 scores from healthy (494 samples) and non-healthy (646 samples) groups. Scores are significantly higher in the healthy group compared to the non-healthy group ($P = 1.6 \times 10^{-43}$; two-sided Mann–Whitney *U* test). The effect size is represented by Cliff's Delta (^{d} = 0.48). The balanced accuracy of the classification is 72.1%. **(b)** GMWI2 scores across five healthy (H^{1} to H^{5}) and three non-healthy cohorts (AS^{4}, ankylosing spondylitis; PD^{6}, Parkinson's disease; PC^{5}, pancreatic cancer). The superscript numbers adjacent to phenotype abbreviations correspond to specific studies detailed in **Supplementary Data 6**. Asterisks (*) indicate significantly higher scores in healthy cohorts compared to non-healthy cohorts ($P < 0.01$, two-sided Mann–Whitney *U* test). Numbers next to each asterisk refer to the healthy cohort compared against each non-healthy condition. Sample sizes for each cohort are shown in parentheses.

We also provide a through interpretation of this data in our new text on **pages 22–23**:

“Demonstration of GMWI2 predictive capability on independent sample sets
*To further confirm GMWI2’s predictive capability for distinguishing between healthy and non-healthy individuals, we compiled an external validation dataset consisting of 1,140 stool metagenome samples from six published studies (**Supplementary Data 6**). This dataset includes samples from healthy individuals and patients diagnosed with ankylosing spondylitis, pancreatic cancer, or Parkinson’s disease. All metagenome samples in this independent and diverse dataset (**Supplementary Data 7**) were classified into either healthy or non-healthy groups, as demonstrated above.*

*Consistent with our findings from the discovery cohort (or training data), GMWI2 scores from stool metagenomes of the healthy validation group (n = 494) were significantly higher than those of the non-healthy validation group (n = 646) ($P = 1.6\times10^{-43}$, Mann–Whitney U test; Cliff’s Delta = 0.48; **Fig. 5a**). The balanced accuracy achieved was 72.1%, which is comparable to the average balanced accuracy of 75.8% observed in our ISV analysis. With magnitude cut-offs of 0.5 and 1.0, balanced accuracy improved to 75.4% and 80.1%, respectively, while still retaining 74.3% and 49.3% of the independent sample sets.*

*To further examine GMWI2 performances across the validation cohorts, we analyzed the eight total cohorts (defined by unique phenotypes from individual studies), spanning four healthy and non-healthy phenotypes. As shown in **Fig. 6b**, four of the five healthy cohorts (H^1–H^4) were found to have significantly higher GMWI2 distributions than all three non-healthy phenotype cohorts ($P < 0.01$, Mann–Whitney U test). Classification accuracies for the healthy cohorts were as follows: 96.3% (130 of 135) for H^1, 91.2% (52 of 57) for H^2, 83.3% (25 of 30) for H^3, 56.8% (21 of 37) for H^4, and 28.1% (66 of 235) for H^5. Alternatively, classification accuracies for the non-healthy cohorts were 90.7% (39 of 43) for pancreatic cancer (PC^5), 81.2% (398 of 490) for Parkinson’s disease (PD^6), and 80.5% (91 of 113) for ankylosing spondylitis (AS^4). Notably, GMWI2 performed well (81.2%) in predicting adverse health in Parkinson’s disease, although stool metagenomes from patients with this neurodegenerative disorder were not part of the original discovery set. Furthermore, despite the relatively poor classification performance in the H^5 cohort (28.1%), the GMWI2 scores in H^5 were significantly higher than those in the PC^5 pancreatic cancer group from the same study. Overall, the robust reproducibility of GMWI2 on an external validation dataset suggests that a generalized disease-associated signature of gut microbiome dysbiosis was effectively captured during dataset integration and index formulation.”*

5. The authors highlight the ability of incorporating magnitude of the GMWI2 scores to improve accuracy, biologically this is a highly rational approach, however, further exploration of the optimal values would be of substantial interest. While such values would be impacted by the nature of the training sets, etc. some attempt should be made to achieve this.

Authors’ Response: Thank you for your suggestion regarding the exploration of an “optimal” GMWI2 cutoff parameter. We wish to emphasize that determining a fixed “optimal” cutoff may not be as straightforward as it seems. The inherent trade-off between higher accuracy (by increasing the magnitude cutoff) and the consequent exclusion of potentially valuable samples is a key consideration in our methodology. This trade-off is an inverse relationship between accuracy and the volume of samples eligible for class prediction, which could limit the broader applicability of the GMWI2 index in diverse gut microbiome datasets.

As detailed in our discussion around **Fig. 3d** in our manuscript, we have provided a comprehensive analysis of these trade-offs. Here, as originally mentioned, the 0.5 and 1.0 cutoffs were merely provided as example discussion points (see this point mentioned on **page 16, lines 119–120**). We strongly believe that the choice of a GMWI2 magnitude cutoff should be flexible and user-defined, tailored to the specific context and acceptable accuracy thresholds of their individual datasets. We have included these points on **pages 29–30, lines 345–356:**

> *"In our analyses in which we incrementally increased the GMWI2 magnitude cutoff, we recognize an inverse relationship between classification accuracy and the volume of samples eligible for class prediction. Therefore, constraining this magnitude cutoff to a single, optimal value may not be universally applicable; instead, the selection of this parameter should be flexible and determined by the user, tailored to the specific context and acceptable accuracy thresholds of their individual datasets. In other words, users can select their desired GMWI2 magnitude cutoff based on their confidence level preference in the predictions. This user-driven approach, which offers flexibility between high confidence in a limited dataset and broader range predictions with lesser confidence, is a distinct advantage of our method over traditional binary-output machine learning techniques. Moreover, our findings thus foster the potential utility of a "reject option"[41,42] for low GMWI2 magnitudes, which can serve as a criterion to redirect relatively uncertain predictions to other screening methods—this concept captures the understanding that certain aspects of health and disease are not fully explainable solely by the gut microbiome."*

6. The authors should be commended on dedicating time to evaluating the impact of individual studies on the robustness of the GMWI2 approach. It is interesting to note the ISV classification accuracy does not appear to correlate with any of the parameters shown. Does this indicate the quality of the definition of health may be a major contributor to the observed results. The authors should include further analysis to explore this phenomenon.

Authors' Response: Thank you for the question. Could the reviewer please clarify what is meant by "parameters shown"? The inter-study validation (ISV) approach demonstrates the substantial variability in classification performance depending on the choice of validation set. (An often overlooked aspect in the literature is the "luck of the draw" in performance on an external validation set.) Specifically, it illustrates the range of classification accuracies one would obtain when using 54 independent validation sets.

Regarding the discussion of the "definition of health", we respectfully maintain our position. The definition of health in this context is predicated on the absence of a clinical diagnosis of disease, which is consistent with our original GMWI work. As the current work represents a continuous refinement of our previous method, we have opted to retain the same definition of "Healthy". However, we acknowledge that this is indeed an intriguing question, albeit one that poses challenges. There is no universal consensus on what "Healthy" truly encompasses, and devising an alternative definition is not a straightforward task, further complicating subsequent analyses. Nevertheless, we have highlighted this as a potential area for future research on **page 31, lines 390–394**:

> *"Last, our definitions of Healthy (i.e., self-reported absence of a disease or disease-related symptoms) and Non-Healthy (i.e., patients with a clinical diagnosis of a disease) are consistent with those used in our previous studies[10,11], as the current work represents a continuous refinement of our previous method. However, we have not investigated how subtle variations in these definitions may impact GMWI2 classification accuracy. Analyzing this aspect is a potential area for future research."*

7. The authors present a small number of case studies to suggest the suitability of the GMWI2 approach. These should be subjected to far greater analysis and specific evidence to demonstrate both the applicability of the GMWI2 approach and the generalisability of the claims of suitability for application to that disease state if they are to be included as evidence to support the suitability of the method in these contexts. For example, the authors suggest "GMWI2 provides more direct relevance to subject phenotype following FMT treatment for IBS" but further evidence is required to conclusively demonstrate this statement across multiple independent datasets. Furthermore, to really demonstrate broad applicability of the GMWI2 index in this context it would be necessary to determining the ability to detect optimal FMT engraftment across broader disease states.

Authors' Response: Thank you for raising this important concern. We would like to clarify that our application of GMWI2 to stool metagenomes from four recent longitudinal gut microbiome studies aimed to demonstrate the tool's *versatility* in tracking longitudinal gut health. This approach was intended to *explore* GMWI2's utility beyond its initial use in case-control scenarios, rather than to reassert its suitability, which we believe has been adequately established in earlier sections of our manuscript. We have clarified this on **page 24, lines 241–247** as follows:

> *"We applied GMWI2 to stool metagenomes obtained from four recently published longitudinal gut microbiome studies. Importantly, these samples were not part of the initial pool of 8069 metagenomes used to train GMWI2. Here, our aim was to illustrate GMWI2's versatility by demonstrating it towards gut microbiome health tracking, thereby extending its applicability beyond the originally intended case-control scenarios. For longitudinal applications, our index for quantitatively monitoring gut health can be likened to using a cholesterol test for evaluating cardiovascular health or a credit score for assessing financial credit history."*

We recognize that the way we initially wrote our interpretation may have extended beyond the specific scope of our demonstration study—thank you for pointing this out. To avoid overgeneralization, we have refined the relevant statement in our manuscript. On **page 24, lines 257–259**, it now reads:

> **Before:** *"Overall, these findings suggest that while α-diversity metrics, such as richness and Shannon diversity, may yield conflicting conclusions, GMWI2 provides more direct relevance to subject phenotype following FMT treatment for IBS."*

> **After:** "*Overall, these findings suggest that while α-diversity metrics, such as richness and Shannon diversity, may yield conflicting conclusions, changes in GMWI2 could serve as a marker of subjects' phenotypes following FMT treatment for IBS.*"

Furthermore, our intention was not to convey any definitive conclusions about FMT. Thus, we removed the following sentence due to this point the reviewer has correctly raised:

> *"In this context, GMWI2 could serve as an effective assessment tool for the dynamic monitoring of clinical symptom relief following FMT."*

Lastly, while we truly acknowledge the significance of "determining the ability to detect optimal FMT engraftment across broader disease states," we believe this objective falls well outside the current scope of our study. However, it undoubtedly represents a compelling direction for future research.

8. As suggested for the first FMT example, the authors should also include samples from multiple diet interventions and antibiotic treatments to demonstrate the broad applicability of the index in this context. These dietary intervention studies are obviously very different in bias and nature to intervention with large complex microbiome communities in FMT.

Authors' Response: We appreciate the reviewer's concern. The reanalysis of existing longitudinal gut microbiome studies in four different contexts was primarily intended to showcase the versatility of GMWI2. Our aim was to demonstrate its applicability beyond the originally intended case-control scenarios by presenting four case studies of longitudinal gut microbiome tracking. These case studies encompassed a study on FMT effects, a diet intervention study, an antibiotic intervention study, and an *in vitro* batch fermentation system of human stool microbiota. We emphasize that the purpose of our analyses was not to draw generalizable conclusions specific to the field of each case study, nor to make cross-case study comparisons, as mentioned by the reviewer. Therefore, including more than one study in each category would have been redundant and unnecessary for the scope of our work. In response to the reviewer's comment, we have made the following modifications:

- removed the following text, which may have been the source of confusion:
  - "Utilizing GMWI2 for new insights from existing longitudinal gut microbiome studies"
  - "Reanalyzing and reinterpreting previously studied datasets can be immensely valuable to gain new insights from existing data."

- added a clarification on **page 24, lines 240–247**:
  - "***Longitudinal gut health tracking in diverse case studies***

    We applied GMWI2 to stool metagenomes obtained from four recently published longitudinal gut microbiome studies. Importantly, these samples were not part of the initial pool of 8069 metagenomes used to train GMWI2. Here, our aim was to illustrate GMWI2's versatility by demonstrating it towards gut microbiome health tracking, thereby extending its applicability beyond the originally intended case-control scenarios. For longitudinal applications, our index for quantitatively monitoring gut health can be

> *likened to using a cholesterol test for evaluating cardiovascular health or a credit score for assessing financial credit history."*

9. While the authors refer to a decrease in gut health associated with the reduced GMWI2 score when analysing stool samples it is important to remember the score is generated from a microbiome analysis of a stool provided from a healthy or diseased individual. There is no measure made of gut health or the actual microbial community within the gut. These compositional changes the model is trained upon could be the result of changes in transit time, stool consistency or many other factors not considered within the metadata. The manuscript should be revised to clearly highlight the association nature of the predictions and the conclusions worded to acknowledge these limitations.

Authors' Response: Thank you for highlighting this important point. We agree and wish to emphasize that GMWI2 scores indicate a statistical association with health status defined as the presence or absence of disease, and not with direct measures of known pathogenic organisms in the gut, metabolic characteristics, serology or inflammatory markers, fecal calprotectin, or other clinical measures of health. Indeed, this is a common limitation in the microbiome field concerning human health. As per your suggestion, we have revised our manuscript to eliminate any implications of causality between GMWI2 scores, the gut microbiome, and human health. We now mention this on **page 30, lines 358–362**:

> *"First and foremost, we emphasize that GMWI2 scores reflect an association with health status, which we define in terms of the presence or absence of disease. It is important to understand that these scores do not imply a causal relationship with direct clinical health measures, including the detection of pathogenic organisms in the gastrointestinal tract, gut motility characteristics, metabolic profiles, serological markers, blood inflammatory markers, or fecal calprotectin levels."*

Regarding the reviewer's comment, *"There is no measure made of gut health or the actual microbial community within the gut"*, we would like to address two points:

> 1. It is true that traditional measures of gut health (such as gut motility characteristics, stool morphology, and fecal calprotectin) are rarely reported in gut microbiome studies, making <u>large-scale</u> correlations with gut microbiome features challenging. This limitation is well recognized in the field.
>
> 2. Contrary to the assertion, <u>GMWI2 directly quantifies taxonomic features from the gut microbial community</u>. This methodology is detailed in the discussion of **Figs. 2a–b**. To clarify, our model is trained on gut microbiome taxonomic profiles, as we have now explicitly state in our newly added text on **page 12, lines 39–41**:
>
> > *"More specifically, the Lasso-penalized logistic regression model utilized 95 microbial taxa with non-zero coefficients for its predictions, derived directly from gut microbiome communities (**Fig. 2b** and **Supplementary Data 3**)."*

Regarding the concern that *“These compositional changes the model is trained upon could be the result of changes in transit time, stool consistency or many other factors not considered within the metadata”* we acknowledge this possibility in individual cases. However, across our study population of over 8000 metagenome samples, we anticipate that these confounding factors are either randomized or minimally impact the overall accuracy of GMWI2. This expectation is based on the scale and diversity of the sample set. We now mention this on **page 31, lines 382–394**:

> *“Seventh, we recognize that compositional shifts between healthy and non-healthy identified by our model might be influenced by variables such as gut transit time, stool consistency, and other factors not captured in our metadata. This is a valid consideration for individual samples. However, in our analysis of over 8000 metagenomic samples, we expect the large sample size and diversity to mitigate the influence of these potential confounders. Our assumption is that such variables are likely to be evenly distributed or have minimal impact on the overall performance and accuracy of the GMWI2 tool, given the breadth and diversity of our dataset. This perspective aligns with epidemiological principles, considering the scale and heterogeneity inherent in our study’s sample population.”*

Minor Comments:
In general the figures are difficult to follow, particularly Figure 1. The authors could consider presenting less text to improve interpretation and data quality.

Authors’ Response: We respect your opinion but respectfully disagree. Fig. 1a includes text that is very similar in length to a typical CONSORT diagram, which is commonly used in clinical literature. We believe this format effectively communicates the essential details of our study design. Could you please specify why you find the flow diagram challenging to follow? Additionally, how would reducing the text in all figures improve the interpretability and quality of the data in the manuscript? It is important to note that this is a manuscript, not a PowerPoint presentation—the audience is expected to engage with the content thoroughly. If there are particular instances of unnecessary jargon that are of high concern, we would appreciate it if you could highlight them for us.

1. In the introduction the authors state "incorporating data from all taxonomic ranks allows microbial signatures across various phylogenetic depths to be captured" - this should be reworded to recognise the critical distinction between taxonomy and phylogeny in this context

Authors’ Response: We appreciate your attention to this detail. Upon reflection, we agree that the term “phylogenetic depths” is not the most appropriate in this context. Therefore, we have revised the wording as follows:

> **before:** *“Additionally, incorporating data from all taxonomic ranks allows microbial signatures across various phylogenetic depths to be captured, and may be critical for optimally predicting host phenotype from microbial communities[14].”*

> **after (page 4, lines 72–74):** *"Additionally, including data across all taxonomic ranks has the potential to uncover more microbial features that accurately predict the host phenotype based on microbial community analysis[14,15]."*

2. The authors should clarify the standardised bioinformatics pipelines does not remove all batch effects, only those associated with the bioinformatics analysis. Collection method, storage, extraction, etc. will still play a role.

Authors' Response: Thank you for this important note! We have added the following text on **page 11, lines 21–25**:

> *"Nevertheless, although the consensus preprocessing of metagenomic data effectively reduces one source of batch effects related to bioinformatics analyses, it is important to recognize that this approach does not eliminate potential batch effects arising from experimental and technical procedures across different studies. Such factors include differences in how stool samples were collected, stored, and prepared for metagenomic sequencing."*

3. The authors cite a publication [25] suggesting there are "on-going concerns about the long-term safety and efficacy of FMT in treating patients with chronic disease"; however, the evidence, including the conclusion from the cited study, suggests FMT is safe and efficacious for multiple diseases most notably UC a chronic condition. The authors should consider refocusing this statement.

Authors' Response: We thank you for highlighting this aspect of our discussion. We acknowledge that there have been previous reports characterizing FMT as safe and effective for various diseases, particularly ulcerative colitis (UC)—but FMT is still far from reaching universal consensus in the clinical landscape. Moreover, although multiple studies suggest that FMT is not inherently hazardous and may be efficacious within the bounds of controlled research settings, we acknowledge that this does not constitute a definitive resolution to the issue. But to avoid any potential misinterpretation, we have removed the statement noted above from our manuscript.

Additionally, we recognize the inherent risks associated with transferring fecal microbiota from an asymptomatic but dysbiotic source. The transfer of a microbiome rich in pathogenic or otherwise disruptive organisms can indeed pose a significant threat, even if no clinical disease has been formally diagnosed. Thus, a more nuanced definition of gut health, such as that provided by GMWI2, could be helpful in identifying and mitigating such risks. There is currently a need for effective tools to distinguish "great" donors and samples from merely "good" ones.

Our intent was to emphasize that tools like GMWI2 can play a valuable role in navigating the complex task of selecting appropriate donors and stool samples for FMT. We have made the necessary clarification in our manuscript on **page 21, lines 257–262**, stating:

> *"Overall, these findings suggest that while α-diversity metrics, such as richness and Shannon diversity, may yield conflicting conclusions, changes in GMWI2 could serve as a marker of*

subjects' phenotypes following FMT treatment for IBS. Furthermore, in light of the clinical significance and the complexities involved in donor screening for FMT[24,25], computational tools such as GMWI2 (given its more nuanced definition of gut health) could help in guiding the selection of suitable healthy donors and their stool samples."

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I appreciate the revisions made by the authors; the manuscript has been significantly improved, and most of my concerns have been addressed. However, regarding the translational potential of this model, I remain unconvinced that it can serve as an early warning system.

In my opinion, there are two primary scenarios where an early-warning model would prove useful. First, it would be super helpful to alarm ongoing change from healthy to disease when an individual is still in a healthy state. Second, it would also be valuable to provide early warnings of disease progression, from its initial stages to more advanced ones.

In the first scenario, such a model would ideally be trained using longitudinal data to differentiate between the composition of a healthy gut microbiome that may transition into a diseased state and one that will not. However, this kind of datasets may be very limited. In the second scenario, the model may need to be tailored to specific diseases because the concept of "stages" varies across different diseases.

The model presented in this study aims to distinguish between a healthy state and a microbiome state associated with confirmed disease. However, it's worth noting that this composition may differ from that observed before the onset of the disease in the same individuals."

Reviewer #2 (Remarks to the Author):

The authors have provided a thoroughly improved manuscript that addresses and clarify the concerns raised previously. While I maintain that Figure 1 could be revised to improve clarity this point can be left to the discretion of the authors and editorial staff.

I have no further concerns.

Reviewer #2 (Remarks on code availability):

The documentation is sufficient to replicate the examples with appropriate README files. Installation instructions were suitable and sufficient to allow installation and running of the code.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I appreciate the revisions made by the authors; the manuscript has been significantly improved, and most of my concerns have been addressed. However, regarding the translational potential of this model, I remain unconvinced that it can serve as an early warning system.

In my opinion, there are two primary scenarios where an early-warning model would prove useful. First, it would be super helpful to alarm ongoing change from healthy to disease when an individual is still in a healthy state. Second, it would also be valuable to provide early warnings of disease progression, from its initial stages to more advanced ones.

In the first scenario, such a model would ideally be trained using longitudinal data to differentiate between the composition of a healthy gut microbiome that may transition into a diseased state and one that will not. However, this kind of datasets may be very limited. In the second scenario, the model may need to be tailored to specific diseases because the concept of "stages" varies across different diseases.

The model presented in this study aims to distinguish between a healthy state and a microbiome state associated with confirmed disease. However, it's worth noting that this composition may differ from that observed before the onset of the disease in the same individuals.

Authors' Response: We appreciate the Reviewer's detailed scenarios and would like to address these points, concluding with our manuscript revisions to temper, but not dismiss, the translational potential of our model.

We fully concur with the Reviewer's ideal scenarios for an early-warning model. However, collecting datasets for training machine learning models for each scenario poses significant challenges. For the first scenario, as the Reviewer rightly points out, collecting longitudinal gut microbiome datasets with the aim of identifying when a healthy state transitions to a diseased state is exceedingly limited and costly, both in terms of resources and labor, especially at the necessary large scale (e.g., >10k samples). Additionally, the Reviewer correctly notes that an individual's gut microbiome composition may differ in a pre-diseased state compared to a confirmed-diseased state. While these points underscore the need for longitudinal data to develop a model that can detect pre-diseased gut microbiomes, it does not invalidate the translational potential of our current approach.

While our index is only explicitly trained to distinguish between healthy (i.e., no disease) and confirmed-diseased gut microbiomes, the interpolation between these states is the most practical way to approximate a pre-diseased microbiome state. Therefore, our trained index can implicitly uncover variation across the "gut microbiome health spectrum" and not just at the extremes. Specifically, assuming sufficient prediction quality of our model, an individual's GMWI2 score (the predicted log-odds that a gut microbiome sample originates from a healthy

individual) will decrease as they transition from healthy to pre-diseased to confirmed-diseased states (or increase if transitioning in the reverse direction). As stated in the manuscript, if an individual observes shifts in GMWI2 scores over time, this could "inform lifestyle modifications to prevent mild issues from escalating into severe health conditions or prompt further diagnostic tests."

Figs. 6b and 6c demonstrate this in two longitudinal case studies, where changes in an individual's GMWI2 reflect health or diet perturbations and subsequent recoveries. While these case studies follow perturbations relevant to the gut microbiome, the existence of ideal longitudinal data following transitions between healthy, pre-diseased, and confirmed-diseased states is currently very limited in scale and diversity. We have revised the manuscript to reflect this limitation.

For the second scenario, identifying early disease stages across various diseases is practically very challenging. While early Type 2 diabetes studies on "glucose tolerance impaired" individuals exist, scaling this approach to multiple diseases remains problematic. The Reviewer rightly points out that capturing early disease stages and tailoring early warning signs to individuals has seen limited success due to practical limitations related to cost, timing, and disease specificity. Additionally, only diseases with substantial funding can reliably pursue such comprehensive studies.

Our approach offers a pragmatic alternative. Instead of focusing on individual diseases, we compare gut microbiomes in healthy (i.e., no disease) states to those with clinical diagnoses, aiming to identify a pan-disease signature of the gut microbiome. This broader comparison addresses fundamental questions such as "What microbial indicators reflect a healthy gut?" and "How can we objectively measure gut health at scale?".

By developing an objective and quantitative index from an extensive corpus of gut microbiome data, we aim to advance wellness, precision nutrition, probiotics, healthy aging efforts, and beyond. Unlike existing indices that are disease-specific, our index spans multiple diseases, emphasizing a general "healthy" signature over specific conditions. This broader focus offers insights beyond the interests of clinical specialists in particular diseases.

We have revised the manuscript to reflect these points and to moderate any excessive claims regarding the model's translational potential. In response to the Reviewer's concern, we have also revised the manuscript to clarify (or remove) mentions of "early screening," "early change," and "early warning system" to better align with the understanding mentioned above.

Reviewer #2 (Remarks to the Author):

The authors have provided a thoroughly improved manuscript that addresses and clarify the concerns raised previously. While I maintain that Figure 1 could be revised to improve clarity this point can be left to the discretion of the authors and editorial staff.

I have no further concerns.

Reviewer #2 (Remarks on code availability):

The documentation is sufficient to replicate the examples with appropriate README files. Installation instructions were suitable and sufficient to allow installation and running of the code.

Authors' Response: We sincerely appreciate the reviewer's time and effort in thoroughly reviewing our paper. We are pleased that the manuscript is now acceptable for publication.