nature communications

Article

https://doi.org/10.1038/s41467-024-51651-9

Gut Microbiome Wellness Index 2 enhances health status prediction from gut microbiome taxonomic profiles

Received: 25 September 2023

Accepted: 9 August 2024

Published online: 28 August 2024

Check for updates

Daniel Chang[1,13], Vinod K. Gupta[2,13], Benjamin Hur[2], Sergio Cobo-López[3], Kevin Y. Cunningham[4], Nam Soo Han[5], Insuk Lee[6], Vanessa L. Kronzer[7], Levi M. Teigen[8], Lioudmila V. Karnatovskaia[9], Erin E. Longbrake[10], John M. Davis III[7], Heidi Nelson[11] & Jaeyun Sung[2,7,12]

Recent advancements in translational gut microbiome research have revealed its crucial role in shaping predictive healthcare applications. Herein, we introduce the Gut Microbiome Wellness Index 2 (GMWI2), an enhanced version of our original GMWI prototype, designed as a standardized disease-agnostic health status indicator based on gut microbiome taxonomic profiles. Our analysis involves pooling existing 8069 stool shotgun metagenomes from 54 published studies across a global demographic landscape (spanning 26 countries and six continents) to identify gut taxonomic signals linked to disease presence or absence. GMWI2 achieves a cross-validation balanced accuracy of 80% in distinguishing healthy (no disease) from non-healthy (diseased) individuals and surpasses 90% accuracy for samples with higher confidence (i.e., outside the "reject option"). This performance exceeds that of the original GMWI model and traditional species-level α-diversity indices, indicating a more robust gut microbiome signature for differentiating between healthy and non-healthy phenotypes across multiple diseases. When assessed through inter-study validation and external validation cohorts, GMWI2 maintains an average accuracy of nearly 75%. Furthermore, by reevaluating previously published datasets, GMWI2 offers new insights into the effects of diet, antibiotic exposure, and fecal microbiota transplantation on gut health. Available as an open-source command-line tool, GMWI2 represents a timely, pivotal resource for evaluating health using an individual's unique gut microbial composition.

Recent landmark studies have unveiled profound links between the gut microbiome and a variety of complex, chronic diseases[1–9]. Despite these discoveries, how can we tell if a person has dysbiosis? How can we effectively harness unique microbial signatures to quantitatively track our health? These critical questions stand at the forefront of utilizing the gut microbiome as a precise marker for health and wellness.

The potential of the gut microbiome as a marker for deciphering complex, chronic diseases has captivated the scientific community—in response, we recently developed the Gut Microbiome Wellness Index (GMWI) [previously called the Gut Microbiome Health Index (GMHI)][10]. GMWI is a first-of-its-kind stool metagenome-based indicator for assessing health by determining the likelihood of an individual harboring a clinically diagnosed disease solely from their gut microbiome

A full list of affiliations appears at the end of the paper. e-mail: Sung.Jaeyun@mayo.edu

composition, irrespective of the specific disease type[10,11]. This disease-agnostic index was derived from a comprehensive analysis of a pooled dataset comprising 4347 stool shotgun metagenomes from 34 independent studies. GMWI is a logarithmic ratio of the collective abundances—a term encompassing species-level relative abundances and multiple α-diversity metrics—of health- and disease-associated gut microbial species. Evaluating on the pooled dataset, GMWI exhibited a balanced accuracy (i.e., average of the proportions of healthy and non-healthy samples that were correctly classified) of 69.7% in predicting the presence of clinically diagnosed disease. Specifically, the correct classification rates for healthy (disease-free) individuals and those with non-healthy (diseased) conditions were 75.6% and 63.8%, respectively. Moreover, GMWI achieved a balanced accuracy of 73.7% in a validation cohort of 679 stool metagenomes, with the correct classification rates for the healthy and non-healthy subsets being 77.1% (91 out of 118) and 70.2% (394 out of 561), respectively. Since its original publication in 2020, GMWI has been utilized in studies investigating the impact of environmental[12] and genetic/socioeconomic[13] factors on the human gut microbiome, as well as in identifying a 'Longevous Gut Microbiota Signature' species set[13].

Despite the promise of our original GMWI prototype, there are limitations that impede its general applicability. Firstly, GMWI correctly classifies healthy stool metagenomes at a higher success rate than non-healthy ones. This bias may stem from the prevalence-based strategy used to identify health-associated and disease-associated species, which was a fundamental component of the GMWI model. As the non-healthy group encompasses patients with different diseases, this group is inherently heterogeneous; in turn, a prevalence-based strategy may miss subtle taxonomic signatures that are only represented in subsets of non-healthy populations (e.g., cohorts with a specific disease). Secondly, our existing model assigns equal weight to each species without considering potential variances in the importance of individual species. To improve classification accuracy and general applicability, a refined weighting system that accounts for varying strengths of association to host phenotype is needed. Additionally, including gut microbial information from all taxonomic ranks could uncover more features that accurately predict host phenotypes[14,15]. In this study, we present GMWI2, an advanced iteration of the original GMWI that addresses the above limitations and significantly improves classification accuracy in distinguishing between healthy and non-healthy phenotypes.

Results

Pooled analysis of stool metagenomes across health and disease phenotypes

As in our previous work[10], we define "healthy" subjects as those without reported diseases or abnormal body weight conditions (i.e., classified as underweight, overweight, or obese based on reported BMI), whereas "non-healthy" subjects are those confirmed to have a clinical diagnosis of any disease. (Retaining the same definitions for "healthy" and "non-healthy" ensures that the current work represents a continuous refinement of our original GMWI method.) We conducted a pooled analysis of existing 8069 stool shotgun metagenomes (5547 from healthy individuals and 2522 from non-healthy individuals) sourced from 54 independently published studies spanning 26 countries and six continents (Fig. 1a, Table 1, and Supplementary Data 1). These pooled metagenomes are from individuals with one of twelve different health and disease phenotypes (Fig. 1a; healthy, ankylosing spondylitis, atherosclerotic cardiovascular disease, colorectal cancer, Crohn's disease, Graves' disease, liver cirrhosis, multiple sclerosis, nonalcoholic fatty liver disease (or also known as metabolic dysfunction-associated steatotic liver disease [MASLD]), rheumatoid arthritis, type 2 diabetes, and ulcerative colitis) from diverse geographies, ethnicities/races, cultures, and balanced sex representation (Fig. 1b). (Our study and sample selection criteria can be found in the "Methods" section. We provide all subjects' phenotype, age, sex, BMI, and geography [as provided in their respective original study] in Supplementary Data 2.) This substantial increase in sample size, nearly doubling the number of metagenomes included in our previous study, is one notable improvement in GMWI2. Additionally, GMWI2 uses MetaPhlAn3[16] instead of MetaPhlAn2[17] for taxonomic profiling, leveraging an extensively expanded marker database for a more comprehensive and accurate characterization of microbial taxa ("Methods" section).

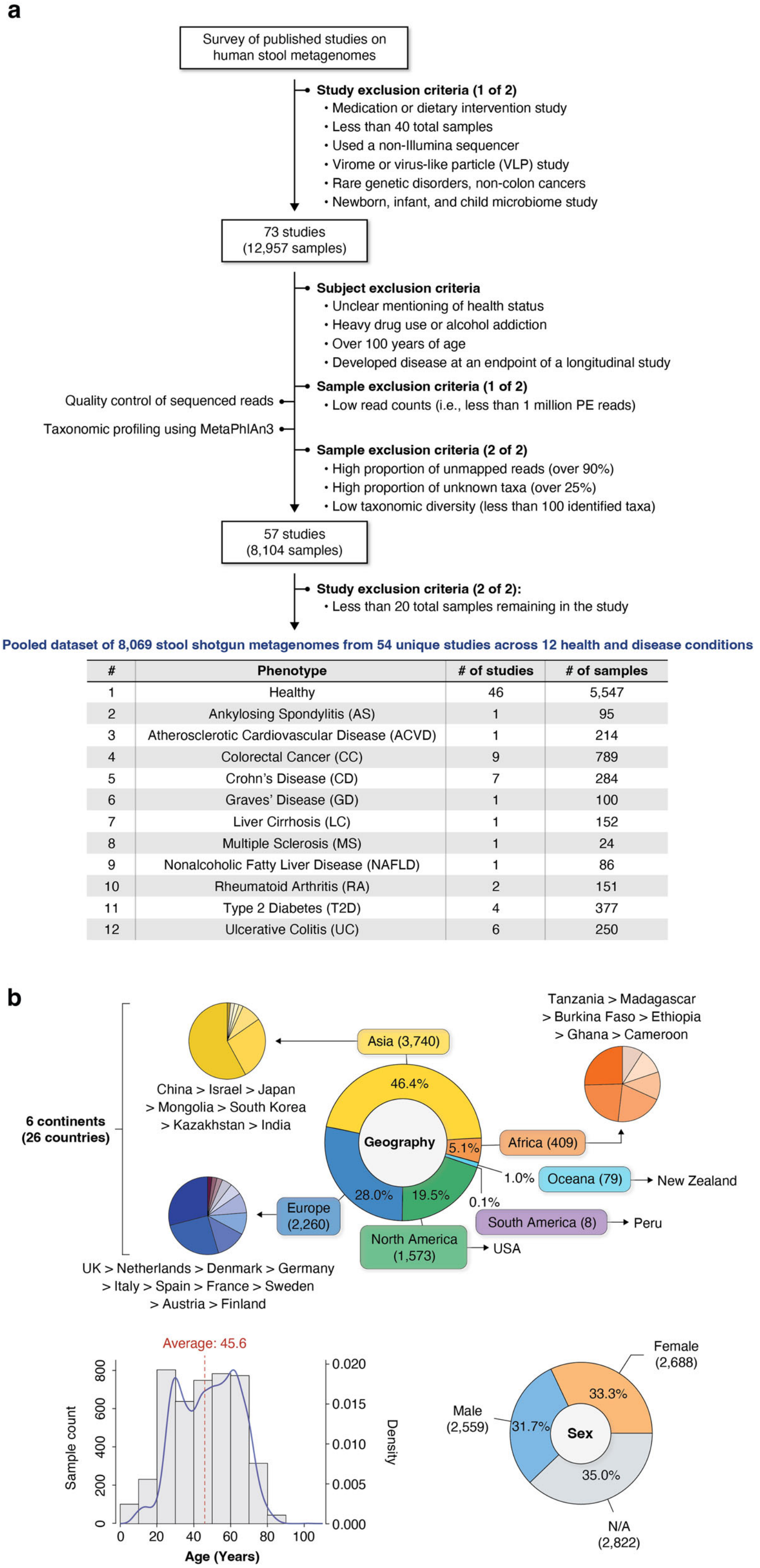
All metagenomes underwent uniform reprocessing using an identical bioinformatics pipeline, as described in the "Methods" section. Such practice not only mitigates batch effects[18,19], but also bolsters the identification of health- and disease-related gut taxonomic signatures despite the presence of potentially strong confounding factors. Indeed, this is supported by principal component analysis (PCA), where, despite the samples originating from varying sources and conditions, the healthy and non-healthy groups display significantly distinct gut microbiome profiles (Adonis $R^2 = 1.2\%$, $P = 0.001$, PERMANOVA; Fig. 2a). 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 cannot entirely 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.

Implementing Lasso-penalized logistic regression in GMWI2

For the classification task of distinguishing between healthy and non-healthy groups, GMWI2 uses a Lasso-penalized logistic regression model instead of the log-ratio equation utilized in the original GMWI. Hence, GMWI2 essentially uses linear regression for its predictions, resembling polygenetic risk score models in statistical genetics[20,21]. The model was trained on gut microbiome taxonomic profiles (derived from the aforementioned pooled dataset of 8069 stool shotgun metagenomes) spanning all measurable taxonomic ranks to model disease likelihood as a linear function of microbial taxon (i.e., clade) presence or absence. Specifically, the GMWI2 score for an individual sample is defined as the predicted log odds (logit) of the sample originating from a healthy, non-diseased individual. A more comprehensive explanation of how GMWI2 uses Lasso-penalized logistic regression to estimate disease likelihood is detailed in "Methods" section.

The original GMWI approach utilized a prevalence-based strategy to identify health- and disease-associated microbial species. Our current method learns variable feature importances, obviating the need for manual species identification. More specifically, the Lasso-penalized logistic regression model utilized 95 microbial taxa with non-zero coefficients for its predictions, derived directly from the gut microbiome profiles (Fig. 2b and Supplementary Data 3). Interestingly, the majority of taxa characterized by positive and negative coefficients exhibited a higher relative abundance in the healthy and non-healthy groups, respectively (Supplementary Data 4). These identified taxa included 1 class, 3 orders, 4 families, 19 genera, and 68 species. Notably, the coefficient values varied between −0.68 and 0.54, ensuring that each taxon contributes differently to the GMWI2 score according to its relative association strength. This presents a shift from our previous GMWI log-ratio model where equal weight was assigned to each species.

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



#	Phenotype	# of studies	# of samples
1	Healthy	46	5,547
2	Ankylosing Spondylitis (AS)	1	95
3	Atherosclerotic Cardiovascular Disease (ACVD)	1	214
4	Colorectal Cancer (CC)	9	789
5	Crohn's Disease (CD)	7	284
6	Graves' Disease (GD)	1	100
7	Liver Cirrhosis (LC)	1	152
8	Multiple Sclerosis (MS)	1	24
9	Nonalcoholic Fatty Liver Disease (NAFLD)	1	86
10	Rheumatoid Arthritis (RA)	2	151
11	Type 2 Diabetes (T2D)	4	377
12	Ulcerative Colitis (UC)	6	250

Fig. 1 | Conducting a pooled analysis of stool metagenomes across multiple health and disease conditions from a diverse global representation. a A survey was conducted in PubMed and Google Scholar to search for published studies with publicly available human stool shotgun metagenome (gut microbiome) samples from healthy (disease-free) and non-healthy (diseased) individuals. The initial collection of stool metagenomes consisted of 12957 samples from 73 independent studies. All raw metagenome samples (.fastq files) were downloaded and reprocessed uniformly using identical bioinformatics methods. After quality control of sequenced reads, taxonomic profiling was performed using MetaPhlAn3. Studies and samples were removed based on several exclusion criteria. Finally, a total of 8069 samples (5547 and 2522 metagenomes from healthy and non-healthy individuals, respectively) from 54 studies ranging across healthy and 11 non-healthy phenotypes were assembled into a pooled metagenome dataset for downstream analyses. **b** Demographic summary of the study subjects whose metagenome samples were included in the pooled dataset. Subject demographics, as reported in the original studies, include country of origin (*n* = 8069), age (*n* = 4670), and sex (*n* = 5247).

Table 1 | Human stool shotgun metagenome datasets used in this study

Author (Last name)	Publication year	Total from study (*n*)	Healthy (*n*)	Non-healthy (*n*)	Disease (*n*)[a]	Sequencing platform	Geography (Country)
Ananthakrishnan	2017	64	0	64	CD (24), UC (40)	Illumina NextSeq 500	United States
Ang	2021	22	22	0	–	Illumina NovaSeq 6000	United States
Asnicar	2021	568	568	0	–	Illumina NovaSeq 6000	United Kingdom/United States
Backhed	2015	100	100	0	–	Illumina HiSeq 2000	Denmark
Costea	2017	169	169	0	–	Illumina HiSeq 2000	Germany/Kazakhstan
D'Souza	2021	128	128	0	–	Illumina NextSeq 500	Netherlands
Davies	2020	44	0	44	T2D (44)	Illumina HiSeq 2000	New Zealand
De Filippis	2019	99	99	0	–	Illumina HiSeq 1500/Illumina NextSeq 500	Italy
Dhakan	2019	47	47	0	–	Illumina NextSeq 500	India
Feng	2015	46	0	46	CRC (46)	Illumina HiSeq 2000	Austria
Franzosa	2018	213	56	157	CD (84), UC (73)	Illumina HiSeq 2000	Netherlands/United States
Gu	2017	94	0	94	T2D (94)	Illumina HiSeq 2500	China
Gupta	2020	49	0	49	RA (49)	Illumina HiSeq 4000	United States
He	2017	86	40	46	CD (46)	Illumina HiSeq 2000	China
Huttenhower; Lloyd-Price[b]	2012; 2017	507	507	0	–	Illumina HiSeq 2000/Illumina Genome Analyzer II	United States
Jacobson	2021	82	82	0	–	Illumina NovaSeq 6000	Burkina Faso
Jie	2017	322	108	214	ACVD (214)	Illumina HiSeq 2000	China
Karlsson	2013	53	0	53	T2D (53)	Illumina HiSeq 2000	Sweden
Kim	2021	61	61	0	–	Illumina HiSeq 4000	South Korea
Le Chatelier	2013	88	88	0	–	Illumina HiSeq 2000/Illumina Genome Analyzer II/Illumina Genome Analyzer IIx	Denmark
Liu	2016	110	110	0	–	Illumina HiSeq 4000	China/Mongolia
Lloyd-Price	2019	86	25	61	CD (39), UC (22)	Illumina HiSeq 2000	United States
Lokmer	2019	37	37	0	–	Illumina HiSeq 2000	Cameroon
Loomba	2017	86	0	86	NAFLD (86)	Illumina HiSeq 2500	United States
Mehta	2018	301	301	0	–	Illumina HiSeq 2000	United States
Nielsen	2014	159	82	77	CD (12), UC (65)	Illumina HiSeq 2000/Illumina Genome Analyzer II/Illumina Genome Analyzer IIx	Denmark/Spain
Obregon-Tito	2015	20	20	0	–	Illumina HiSeq 2500	Peru/United States
Pasolli	2019	142	142	0	–	Illumina HiSeq 2500	Ethiopia/Madagascar
Qi	2019	43	43	0	–	Illumina HiSeq 2500	China
Qin	2012	369	183	186	T2D (186)	Illumina Genome Analyzer II	China
Qin	2014	287	135	152	LC (152)	Illumina HiSeq 2000	China
Rettedal	2021	35	35	0	–	Illumina HiSeq 2500	New Zealand
Roager	2019	50	50	0	–	Illumina HiSeq 2000	Denmark
Schirmer	2016	385	385	0	–	Illumina HiSeq 2000	Netherlands
Schirmer	2018	83	18	65	CD (39), UC (26)	Illumina HiSeq 2000	United States
Smits	2017	38	38	0	–	Illumina HiSeq 4000	Tanzania
Sun	2021	42	42	0	–	Illumina HiSeq 4000	United States
Tett	2019	110	110	0	–	Illumina HiSeq 2000/Illumina HiSeq 2500	Tanzania/Ghana
Thomas	2019	160	61	99	CRC (99)	Illumina HiSeq 2500	Italy/Japan
Ventura	2019	48	24	24	MS (24)	Illumina HiSeq 4000	United States
Vogtmann	2016	81	30	51	CRC (51)	Illumina HiSeq 2000	United States
Wen	2017	200	105	95	AS (95)	Illumina HiSeq 2000	China
Weng	2019	79	15	64	CD (40), UC (24)	Illumina HiSeq X Ten	China
Wirbel	2019	55	33	22	CRC (22)	Illumina HiSeq 4000	Germany
Xie	2016	130	130	0	–	Illumina HiSeq 2000	United Kingdom

Table 1 (continued) | Human stool shotgun metagenome datasets used in this study

Author (Last name)	Publication year	Total from study (*n*)	Healthy (*n*)	Non-healthy (*n*)	Disease (*n*)[a]	Sequencing platform	Geography (Country)
Yachida	2019	217	0	217	CRC (217)	Illumina HiSeq 2500	Japan
Yang	2020	180	88	92	CRC (92)	Illumina HiSeq X Ten	China
Yang	2021	194	97	97	CRC (97)	Illumina NovaSeq 6000	China
Yassour	2018	42	42	0	–	Illumina HiSeq 2500	Finland
Yu	2015	128	53	75	CRC (75)	Illumina HiSeq 2000	China
Zeevi	2015	900	900	0	–	Illumina HiSeq 2500/Illumina HiSeq 2500/Illumina MiSeq	Israel
Zeller	2014	135	45	90	CRC (90)	Illumina HiSeq 2000	France/Germany
Zhang	2015	163	61	102	RA (102)	Illumina HiSeq 2000	China
Zhu	2021	132	32	100	GD (100)	Illumina HiSeq 4000	China

[a] *ACVD* atherosclerotic cardiovascular disease, *AS* ankylosing spondylitis, *CRC* colorectal cancer, *CD* Crohn's disease, *GD* Graves' disease, *LC* liver cirrhosis, *MS* multiple sclerosis, *NAFLD* nonalcoholic fatty liver disease, *RA* rheumatoid arthritis, *T2D* type 2 diabetes, *UC* ulcerative colitis.

[b] Samples combined from both phases of the Human Microbiome Project (HMP1 and HMP1-II).

Further details on individual studies and their metagenome samples can be found in Supplementary Data 1 and Supplementary Data 2.

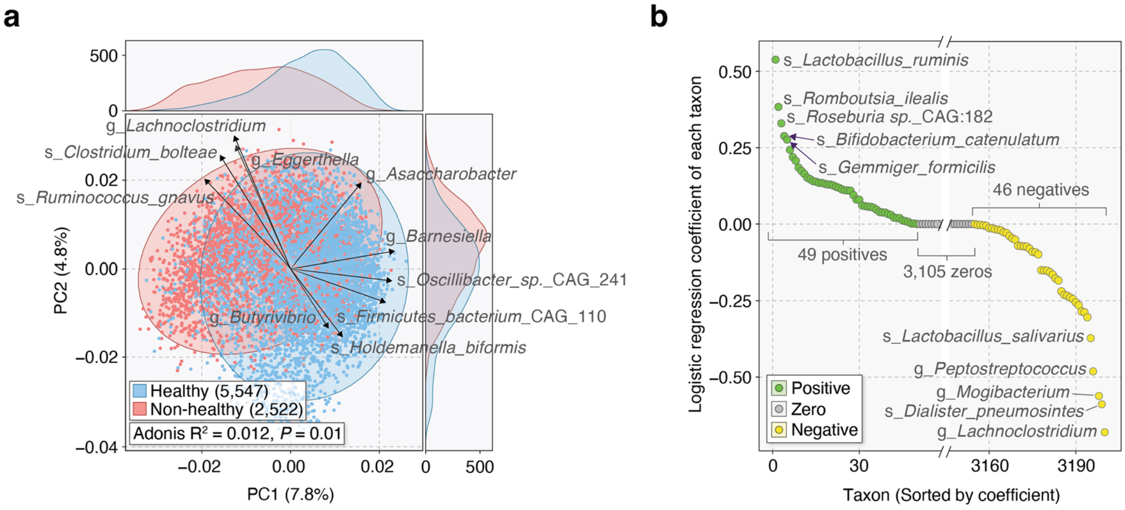


Fig. 2 | Gut microbiome taxonomic profiles of healthy and non-healthy individuals inform a Lasso-penalized logistic regression classification model. **a** Principal component analysis (PCA) of gut microbiome profiles. Significant differences in distributions between healthy (disease-free) (blue, $n = 5547$) and non-healthy (diseased) (red, $n = 2522$) groups were observed ($P < 0.05$, PERMANOVA). Ellipses represent 95% confidence regions. The loading vectors with the top 10 highest PC1 and PC2 magnitudes are shown. **b** Coefficient values for the Lasso-penalized logistic regression model. The model includes 49 taxa with positive coefficients, 3105 taxa with zero coefficients, and 46 taxa with negative coefficients.

Table 1). Given our primary objective of optimizing classification accuracy, we chose to prioritize this aspect, leading us to set aside the multicollinearity concern.

In the following sections, we evaluate GMWI2's proficiency in differentiating healthy from non-healthy individuals. This process can be conceptually structured into four phases:

1. Model training: GMWI2 is trained and evaluated on the full training dataset. This phase utilizes all 8069 samples for computing the logistic regression coefficients (as depicted in Fig. 2b) and determining GMWI2 scores.
2. Cross-validation: GMWI2 undergoes further evaluation through cross-validation (CV) and inter-study validation (ISV) strategies. In contrast to the initial phase, these strategies do not leverage all 8069 samples simultaneously for model training. As a result, the models generated during this phase are intrinsically different from those produced in the first phase. In line with standard cross-validation protocols, the training of the GMWI2 model, including the computation of logistic regression coefficients, is confined strictly to the training partition of each train-test split of the total 8069 samples.
3. Validation on external datasets: The GMWI2 model developed in the first phase is applied to six external datasets to confirm its discriminatory power on independent samples.
4. Demonstration on longitudinal datasets: The GMWI2 model from the first phase is applied to four additional external datasets. These evaluations focus on demonstrating GMWI2's applicability in longitudinal scenarios.

Enhanced classification of healthy and non-healthy gut microbiomes with GMWI2

GMWI2 scores were calculated for metagenomes by applying the learned coefficients in computing the predicted log odds. A positive GMWI2 value classifies the sample as healthy, indicating disease absence; while a negative GMWI2 value classifies it as non-healthy, denoting disease presence. A GMWI2 of 0 implies an equal weighted presence of positive coefficient taxa and negative coefficient taxa, thereby classifying the sample as neither healthy nor non-healthy. When evaluated on the training dataset (8069 samples), GMWI2 demonstrated a balanced accuracy of 79.9% (correct classification rate in healthy: 79.2%, correct classification rate in non-healthy: 80.6%)

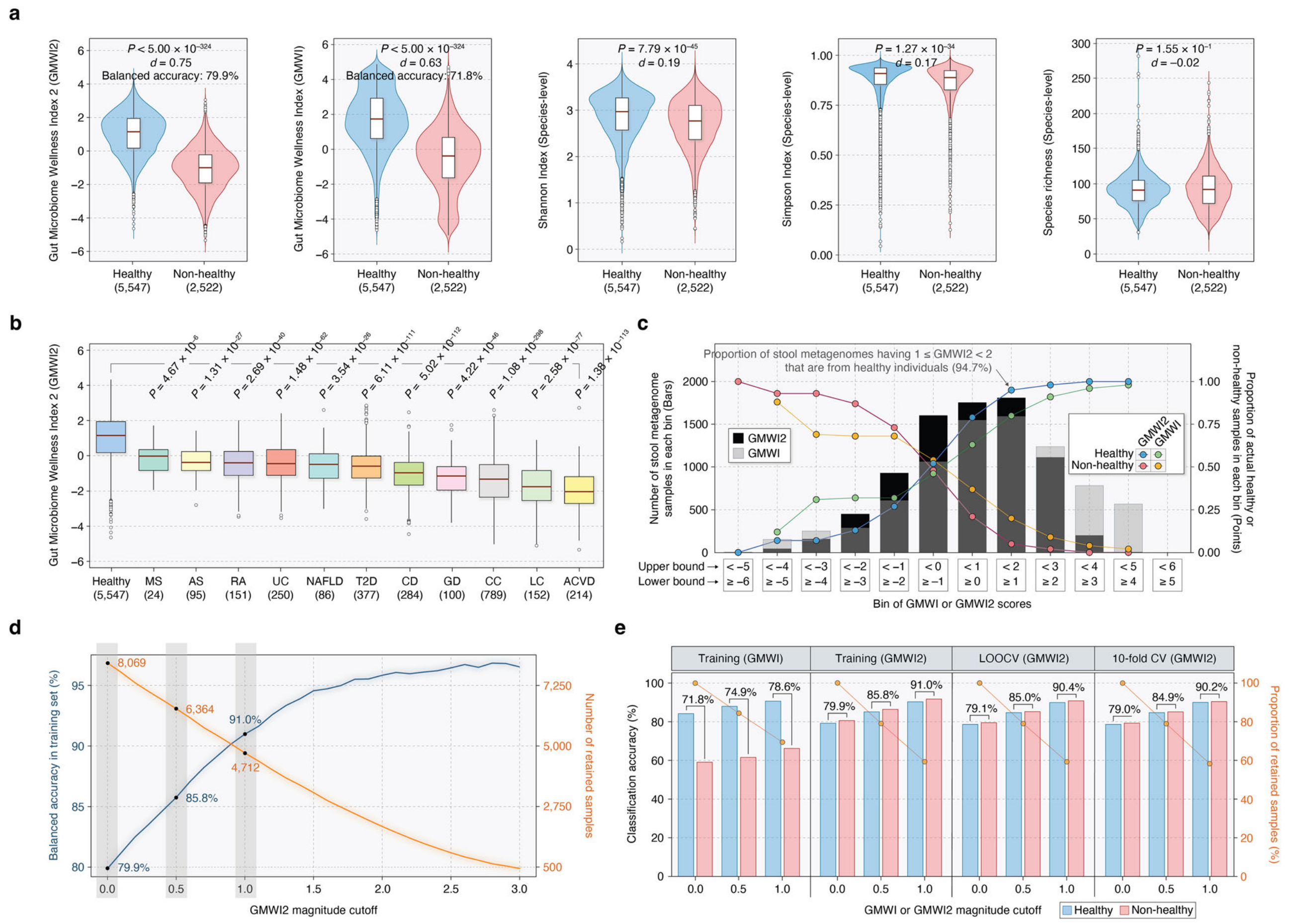


Fig. 3 | Enhanced classification of healthy and non-healthy stool metagenomes using Gut Microbiome Wellness Index 2 (GMWI2). a GMWI2 best stratifies healthy ($n = 5547$) and non-healthy ($n = 2522$) groups compared to GMWI and α-diversity indices (*P*-values from the two-sided Mann–Whitney *U* test; *d*, Cliff's Delta effect size). Balanced accuracies on the training set are shown for GMWI2 and GMWI. **b** The healthy group (blue, far left) exhibits significantly higher GMWI2 scores than all 11 non-healthy phenotypes (*P*-values from the two-sided Mann–Whitney *U* test). Non-healthy phenotypes include multiple sclerosis (MS, $n = 24$), ankylosing spondylitis (AS, $n = 95$), rheumatoid arthritis (RA, $n = 151$), ulcerative colitis (UC, $n = 250$), nonalcoholic fatty liver disease (NAFLD, $n = 86$), type 2 diabetes (T2D, $n = 377$), Crohn's disease (CD, $n = 284$), Graves' disease (GD, $n = 100$), colorectal cancer (CC, $n = 789$), liver cirrhosis (LC, $n = 152$), and atherosclerotic cardiovascular disease (ACVD, $n = 214$). **c** Bins of GMWI2 and GMWI scores (x-axis). The height of the black and gray bars indicate metagenome sample counts in each GMWI2 and GMWI bin, respectively (y-axis, left). Points represent the proportion of samples in each GMWI2 or GMWI bin corresponding to actual healthy and non-healthy individuals (y-axis, right). **d** Increased magnitude cutoffs result in improved classification performance of GMWI2, showing increasing training set balanced accuracy (blue, y-axis, left) at the expense of decreasing retained samples (orange, y-axis, right). **e** Classification performances of GMWI and GMWI2 in distinguishing healthy and non-healthy groups. Accuracies (y-axis, left) are depicted for both groups on the training set, leave-one-out cross-validation (LOOCV), and 10-fold CV, using varying magnitude cutoffs (0, 0.5, 1.0) of GMWI and GMWI2 scores. Balanced accuracies are shown between the blue and pink bars, which represent healthy and non-healthy groups, respectively. Orange points represent the proportion of retained samples (y-axis, right) for the corresponding index magnitude cutoff. For 10-fold CV, repeated random sub-sampling was performed ten times, and the average results are displayed. Standard box-and-whisker plots (i.e., center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers) are used to depict groups of numerical data in (**a**, **b**).

and a Cliff's Delta (*d*) effect size of 0.75, significantly surpassing the balanced accuracy and Cliff's Delta reported by our original GMWI model (71.8%, $d = 0.63$) and traditional species-level α-diversity indices (i.e., Shannon Index, Simpson Index, and richness) (Fig. 3a and Supplementary Data 5). Our results indicate that GMWI2 differentiates between healthy and non-healthy groups much more effectively than GMWI, although both indices were strongly correlated (Pearson's $r = 0.81$; Supplementary Fig. 1). Moreover, we found that the gut microbiomes of healthy individuals exhibit significantly higher GMWI2 scores compared to each of the eleven disease phenotypes (Fig. 3b). Lastly, we observed weak correlations between GMWI2 and clinical/demographic characteristics ($|\text{Spearman's } \rho| < 0.3$; Supplementary Figs. 2a–g), such as age, BMI, fasting blood glucose, blood cholesterol and triglycerides, indicating that these factors do not significantly influence gut microbiome-based classification outcomes.

We subsequently explored whether higher (or more positive) GMWI2 values could indicate enhanced confidence in categorizing stool metagenomes as healthy. Conversely, we examined if lower (or more negative) GMWI2 scores suggest an increased likelihood that a sample could be classified as non-healthy. Indeed, we observed a progressive increase in the proportion of healthy individuals among metagenome samples with increasingly positive GMWI2 scores (Fig. 3c and Supplementary Table 2). Similarly, increasingly negative GMWI2-scores captured larger proportions of the non-healthy subjects. Notably, the proportions of actual healthy and non-healthy samples within the positive and negative bins of GMWI2, respectively, were both higher compared to the same GMWI bins (refer to points in Fig. 3c). This difference in sample distributions between the GMWI2 and GMWI bins underscores GMWI2's improved capability to differentiate between healthy and non-healthy samples.

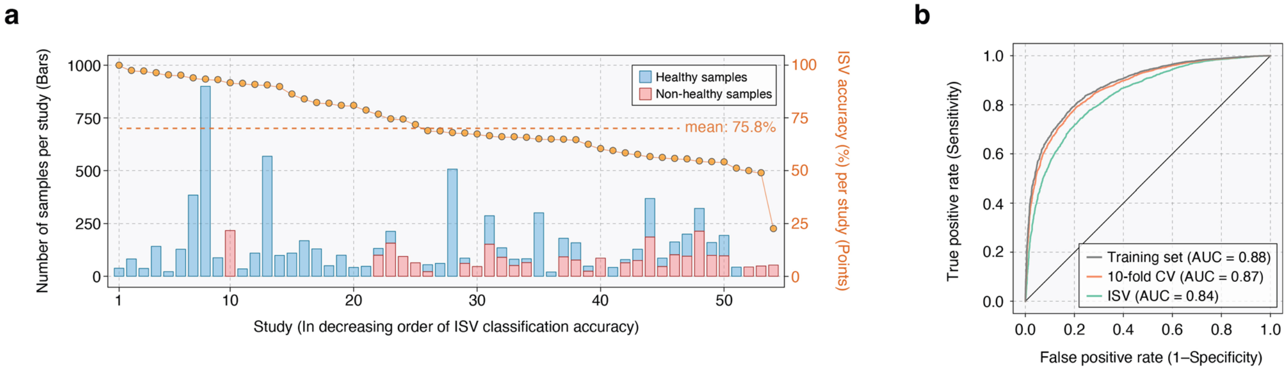


Fig. 4 | Inter-study validation (ISV) shows effective generalization of GMWI2 across diverse study populations. a Classification accuracy on each excluded study in ISV is displayed by gold points (y-axis, right). The studies on the x-axis are rank-ordered based on either accuracy for a single phenotype (healthy or non-healthy) or balanced accuracy in the case of both phenotypes. The stacked bars illustrate the number of healthy (blue) and non-healthy (pink) stool metagenome samples in each study (y-axis, left). **b** Receiver operating characteristic curves for classification performance in distinguishing healthy and non-healthy phenotypes on the training set, 10-fold CV, and ISV.

The results presented in Fig. 3c of our study revealed an interesting 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 observed remarkable improvement in classification performance when considering increasing cutoffs for the magnitude of GMWI2 scores, thereby signifying higher prediction confidence in the retained samples (Supplementary Table 3). For example, when retaining samples with GMWI2 magnitudes equal to or higher than 0.5 (i.e., GMWI2 scores below −0.5 or above +0.5) and 1.0 (i.e., GMWI2 scores below −1.0 or above +1.0), we achieved balanced accuracies of 85.8% and 91.0%, respectively (Fig. 3d). (these cutoffs are examples to illustrate the concept of the GMWI2 magnitude cutoff.) This approach, however, requires excluding samples with GMWI2 magnitudes below these cutoffs, leaving only 6364 (representing 78.9% of the total 8069 samples) and 4712 (58.4% of 8069) samples, respectively. This highlights a significant trade-off: increasing the cutoff improves accuracy but excludes potentially valuable samples from the analysis.

An important observation is that GMWI2 correctly classified healthy and non-healthy stool metagenomes at nearly the same rate (79.2% and 80.6%, respectively) despite imbalanced sample numbers. This contrasts markedly with the original GMWI, which achieved a much higher correct classification rate on healthy samples (Fig. 3e). We also assessed the performance of the GMWI2 model utilizing both leave-one-out cross-validation (LOOCV) and 10-fold cross-validation (10-fold CV) (Fig. 3e). Interestingly, GMWI2 achieved nearly identical balanced accuracies of 79.1% (healthy correct classification rate: 78.6%, non-healthy correct classification rate: 79.5%) and 79.0% (healthy correct classification rate: 78.6%, non-healthy correct classification rate: 79.3%) in LOOCV and 10-fold CV, respectively, nearly matching the performance achieved on the training dataset (79.9%).

Next, we computed classification accuracies using different magnitude cutoffs for the two cross-validation methods (Fig. 3e). Remarkably, GMWI2 achieved a balanced accuracy of 90.4% and 90.2% in LOOCV and 10-fold CV, respectively, on the samples with scores below −1.0 or above +1.0. These balanced accuracies were very close to those observed in the training set (91.0%). In contrast, when applying the same criteria to GMWI (i.e., cutoff of 1.0), the balanced accuracy drops considerably to 78.6%. In all, these results emphasize the notable improvements achieved with GMWI2 over GMWI.

Evaluating the robustness of GMWI2 across study populations of varying sample sizes

Although studies with small sample sizes were excluded from the training set (see study exclusion criteria in Fig. 1a and “Methods” section), in general, it is crucial to validate any classification model on datasets of varying sample sizes[19]. To this end, we conducted inter-study validation (ISV) to assess the impact of batch effects (i.e., technical or biological variations associated with the study population or site characteristics) 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 held-out study[22]. (The excluded study essentially becomes the independent validation [or test] cohort.) An important aspect of ISV is that it can showcase the significant variability in classification performance that can arise depending on the choice of validation set. For our study, it provides a range of classification accuracies achievable when applying GMWI2 across 54 independent validation sets.

Figure 4a specifically displays the performance of GMWI2 across the full range of held-out studies, along with details on their sample sizes. Despite the variation in classification performance across different studies (see gold points indicating ISV classification accuracy per study in Fig. 4a and Supplementary Table 4), 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 (Supplementary Table 4). In all, our analysis revealed no discernible correlation between the model’s predictive performance and the sample size of the held-out datasets.

The classification performances obtained from ISV exhibited minimal disparity compared to the performances achieved by LOOCV and 10-fold CV, which do not consider study boundaries. The small discrepancy between these strategies shows GMWI2’s resilience against batch-related biases, indicating that GMWI2 generalizes effectively across stool metagenomes, regardless of the subjects’ origins. Further evidence of this robustness is demonstrated by the area-under-the-curve (AUC) metrics in the training set, 10-fold CV, and ISV, achieving AUCs of 0.88, 0.87, and 0.84, respectively (Fig. 4b).

Demonstration of GMWI2 predictive capability on independent sample sets

To confirm GMWI2’s predictive capability for distinguishing between healthy and non-healthy individuals, we compiled an external validation dataset consisting of 1140 stool metagenome samples from six

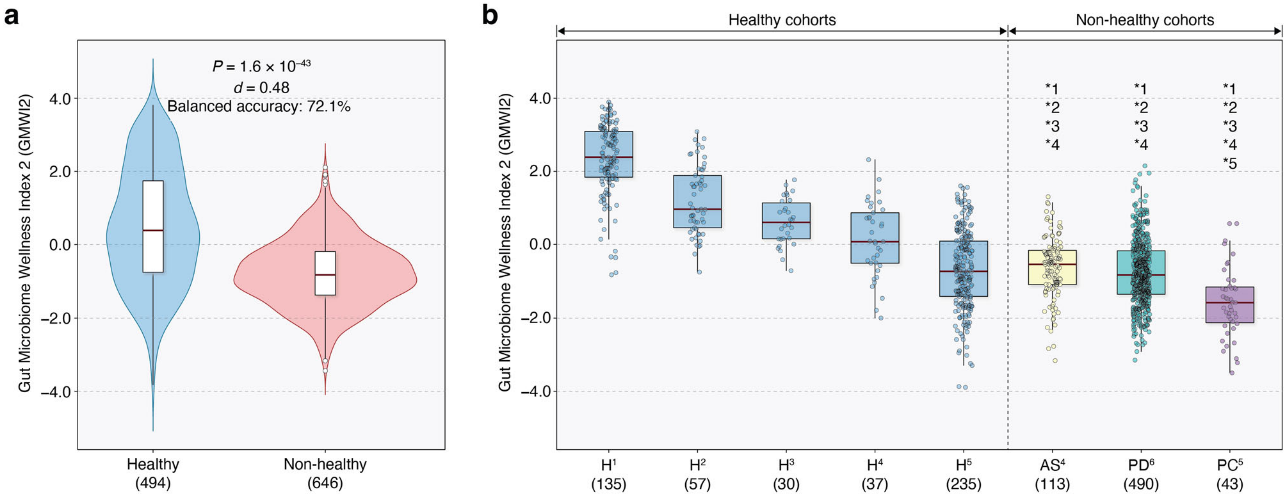


Fig. 5 | GMWI2 performance on healthy and non-healthy external validation cohorts. a GMWI2 scores from healthy (494 samples) and non-healthy (646 samples) groups. Scores are significantly higher in the healthy group compared to the non-healthy group ($P = 1.6 \times 10^{-43}$; two-sided Mann–Whitney *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–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. Asterisk (*) indicates significantly higher score in a healthy cohort compared to the corresponding non-healthy cohort ($P < 0.01$, two-sided Mann–Whitney *U* test. Exact *P*-values provided in Supplementary Data 6). Numbers next to each asterisk refer to the healthy cohort compared against each non-healthy condition. Sample size of each group or cohort are shown in parentheses. Standard box-and-whisker plots (i.e., center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers in (**a**) or individual GMWI2 scores in (**b**)) are used to depict groups of numerical data.

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 validation dataset (Supplementary Data 7) were classified into either healthy or non-healthy groups in the same manner 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 ($n = 646$) ($P = 1.6 \times 10^{-43}$, two-sided 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 cutoffs of 0.5 and 1.0, the balanced accuracy improved to 75.4% and 80.1%, respectively, while still retaining 74.3% and 49.3% of the samples.

To further examine GMWI2 performance on the external validation data, we analyzed the eight total cohorts (defined by unique phenotype per study), spanning five healthy and three non-healthy phenotypes. As shown in Fig. 5b, 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 ($P < 0.01$, two-sided Mann–Whitney *U* test). Classification accuracies for the five 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 three 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 across multiple diseases was effectively captured during dataset integration and index formulation.

Gut health tracking in longitudinal 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 vs. control scenarios. Our index for quantitatively monitoring gut health can be likened to using a cholesterol and glucose test for evaluating cardiovascular and metabolic health over time.

Using data from the first study[23], we analyzed stool metagenomes from 22 individuals with irritable bowel syndrome (IBS) before and six months after receiving fecal microbiota transplantation (FMT) from two healthy donors. Among the participants, 14 reported symptom relief after FMT ("Effect" group), while 8 did not experience symptom relief ("No Effect" group) despite both groups demonstrating a significant increase in species richness at six months following FMT ($P < 0.05$, one-sided Wilcoxon signed-rank test; Supplementary Fig. 3). However, only the individuals in the "Effect" group exhibited a significant increase in GMWI2 ($P < 0.05$; Fig. 6a and Supplementary Table 5). Likewise, an increase in the species-level Shannon Index was observed only in the "Effect" group ($P < 0.05$; Supplementary Fig. 4). 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) may be able to help guide the selection of suitable healthy donors and their stool samples.

In the second study[26], we investigated the effects of diet. We calculated GMWI2 for stool metagenomes obtained from 30 healthy volunteers before and during a dietary intervention. Three groups of participants were studied: Vegan (self-reported vegans who resumed

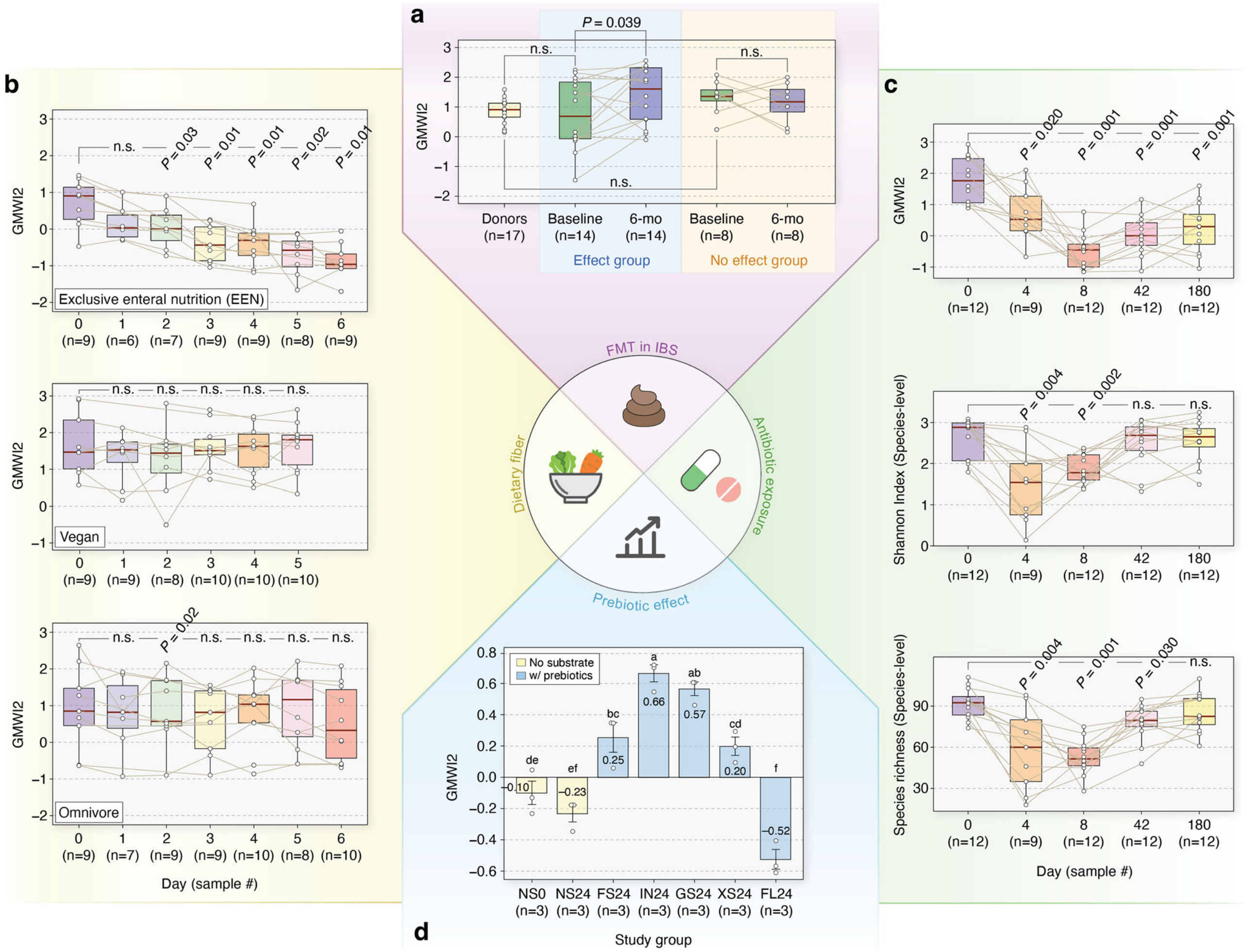


Fig. 6 | Reanalysis of existing longitudinal gut microbiome studies with GMWI2. a Changes in GMWI2 in patients with irritable bowel syndrome observed six months (6-mo) after undergoing fecal microbiota transplantation. Only subjects experiencing symptom relief ("Effect" group) displayed a significant increase in GMWI2 ($P = 0.039$, one-sided Wilcoxon signed-rank test). *n*, number of FMT donor samples (17 total samples from two healthy donors) or number of FMT recipients. **b** GMWI2 scores for dietary groups (EEN, Vegan, and Omnivore) at baseline and at the first 5–6 days of dietary intervention. The EEN group showed significant changes in GMWI2, with values significantly decreased by day 2 and thereafter ($P < 0.05$, two-sided Wilcoxon signed-rank test). No significant change in GMWI2 was observed for the Omnivore and Vegan groups compared to baseline. n, number of unique individuals who each provided a stool sample per time point. **c** GMWI2, Shannon Index, and species richness before and after antibiotic intervention. Despite recovery in Shannon Index and species richness at day 42 and day 180, respectively, GMWI2 remained significantly lower compared to day 0, suggesting incomplete gut microbiome recovery even after ~6 months ($P < 0.05$, two-sided Wilcoxon signed-rank test). *n*, number of unique individuals who each provided a stool sample per time point. **d** GMWI2 of gut microbial communities after 24-h in vitro fecal fermentation with five different prebiotic oligosaccharides. The experiment was conducted in triplicates for each study group. The height of the bars represents the mean GMWI2 (numbers inside the solid bars), and error bars indicate the standard deviation from the mean. Points represent individual triplicate samples. Different small letters above the bars denote groups with significant differences in GMWI2 as determined by Tukey's HSD test ($P < 0.05$). Control groups: NS0, no substrate addition at 0 h; NS24, no substrate for 24 h. Prebiotic groups: FS24 fructooligosaccharide, IN24 inulin, GS24 galactooligosaccharide, XS24 xylooligosaccharide, FL24 2′-fucosyllactose. Standard box-and-whisker plots (i.e., center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, individual GMWI2 scores or α-diversity values) are used to depict groups of numerical data in (**a**–**c**).

their regular diet), Omnivore (participants who consumed a standard diet of both animal and plant origin), and Exclusive Enteral Nutrition (EEN) (participants with an omnivorous diet who went on to consume a synthetic, fiber-free diet for the duration of the study). Stool samples were collected at baseline and each day during the dietary intervention. We observed that the GMWI2 scores for both the vegan and omnivore subjects remained relatively stable throughout the intervention period of five to six days (Fig. 6b). However, GMWI2 for the EEN group significantly decreased relative to baseline by the second day and onwards ($P < 0.05$, two-sided Wilcoxon signed-rank test; Fig. 6b and Supplementary Table 6) while α-diversities did not significantly change across the groups (Supplementary Fig. 5). These results suggest that the removal of dietary fiber may lead to a rapid decrease in overall gut health, an early change detected solely by GMWI2 and not by α-diversity metrics. Overall, our findings strengthen the evidence for the well-established benefits of dietary fiber on health[27–29].

For the third study[30], we calculated GMWI2 for stool metagenomes from twelve healthy young adults who underwent a 4-day exposure with broad-spectrum antibiotics (meropenem, gentamicin, and vancomycin). Here, stool samples were collected before the exposure, and then again at 4, 8, 42, and 180 days post-intervention. While species-level α-diversity measures (Shannon Index and richness) indicated that the gut microbiome may have recovered somewhat by day 42 or 180, GMWI2 did not demonstrate any recovery trend even by day 180 (Fig. 6c and Supplementary Table 7). These findings reflect deleterious post-intervention taxonomic shifts originally noted by Palleja et al., such as the rise in previously undetectable *Clostridium* *spp*., and the disappearance of probiotic members of *Bifidobacterium* and butyrate producers *Coprococcus eutactus* and *Eubacterium ventriosum*. Our results therefore offer a novel perspective on the long-term impact of short-term broad-spectrum antibiotic intervention on gut microbiota and suggest that GMWI2 could be a valuable tool for assessing gut microbiome recovery following an acute illness.

In the final study[31], we examined the effect of various oligosaccharides on gut microbial communities. In this study, Lee et al. used GMWI to assess the prebiotic effect of oligosaccharides, with broader implications for designing personalized diets based on their impact on gut microbiome wellness. Herein, 19 healthy adult volunteers (14 men and 5 women) provided fecal samples, which were then combined and well-mixed. Then, fructooligosaccharides (FOS), galactooligosaccharides (GOS), xylooligosaccharides (XOS), inulin (IN), and 2′-fucosyllactose (2FL) were separately mixed with portions of the homogenized fecal samples in a 24-h in vitro anaerobic batch fecal fermentation system. Two control groups were also included: one without substrate addition at 0 h (NS0) and another without substrate addition for 24 h (NS24). The experiment was conducted in triplicates for each of the seven study groups.

GMWI2 was calculated for all fecal samples (Fig. 6d and Supplementary Table 8), thereby replicating the original study with our new index. Consistent with previous findings, the NS24 group exhibited a lower average GMWI2 than the NS0 group, indicating a less healthy and more disease-associated state. Notably, the addition of the three prebiotics (FOS, IN, and GOS) resulted in significantly higher GMWI2 compared to NS0 ($P < 0.05$, Tukey's HSD test). Also, these same three prebiotics, along with XOS, led to significantly higher GMWI2 relative to NS24 ($P < 0.05$). However, unlike the GMWI2 results, traditional α-diversity metrics (Shannon Index, species richness, species evenness, and inverse Simpson's Index) were reported to have significantly lower values in all prebiotic treatment groups compared to the NS0 group ($P < 0.05$)[31]. Therefore, at least in the in vitro fermentation setting, intake of these four prebiotics could potentially stimulate the growth of gut microbial species associated with healthy conditions, an effect observed solely by using GMWI2.

Discussion

Recent research into the human gut microbiome has highlighted its potential to inform the development of innovative tools for predictive healthcare[32–37]. In this regard, we introduce GMWI2, a robust predictor of health status based on gut microbiome taxonomic profiles that display significant technological advances compared to its prototype (GMWI). Our extensive multi-study analysis, pooling 8069 stool shotgun metagenomes from 54 published studies, encompasses a diverse range of demographics from 26 countries across six continents to identify the biological signals linking gut taxonomies to human health. Delivering a cross-validation balanced accuracy of approximately 90% for higher confidence samples, GMWI2 establishes its strong reliability as a classifier that distinguishes between healthy and non-healthy phenotypes. Furthermore, by revisiting and reinterpreting data from previously published datasets, GMWI2 can offer novel perspectives even for the established understanding of the impact of dietary influences, antibiotic exposure, and FMT on the gut microbiome. Lastly, this study highlights the importance of extensive data sharing in fostering robust machine learning applications, and in demonstrating resilience to batch effects and biases[22,38–40].

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 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.

Our study, while providing insights into the predictive capabilities of the gut microbiome, has some limitations that need to be acknowledged. 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 (nor are they intended to replace) direct clinical health measures, such as the detection of pathogenic organisms in the gastrointestinal tract, gut motility characteristics, metabolic profiles, serological markers, blood inflammatory markers, or fecal calprotectin levels. Second, the model could benefit from the inclusion of more intricate microbiome features such as species growth rates, strain details, and functional potential. Incorporating these important factors may improve predictive accuracy and offer a richer perspective on the intricate mechanisms tying the gut microbiome to overall human health. Third, we made concerted efforts to ensure that our pooled stool metagenomic dataset exhibits a diverse representation of geographies, races, and cultures. Nevertheless, future work should emphasize wider participant inclusion, especially from underrepresented areas and ethnicities, to truly globalize gut microbiome research. Additionally, loosening our selection criteria will allow us to incorporate metagenomes from a broader range of disease phenotypes (like neurodegenerative and psychiatric disorders) and reach even more diverse demographics. Such expansion could enhance the model's generalizability across different populations. Fourth, although we utilized taxonomic information down to the species level, there's a potential missed opportunity in not focusing on microbial strains, which often bear more clinical significance. While our method surpasses the genus-level limitations of 16S rRNA gene amplicon sequencing, it doesn't account for the variability among strains of the same species. 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 diseases necessarily harbors invasive pathogens. Sixth, we recognize that the compositional shifts between healthy and non-healthy identified by our model might be influenced by variables such as transit time, stool consistency, and other factors not captured in our meta-data. This is a valid consideration for individual samples. However, in our analysis of over 8000 metagenome samples, our assumption is that such variables are likely to be evenly (randomly) distributed or have minimal impact on the overall performance of the GMWI2 tool, given the breadth and reasonable diversity inherent in our study's sample population. 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.

In regard to its translational potential, GMWI2 is designed to offer a novel method for dynamically monitoring an individual's health in a semi-real-time manner through the analysis of gut microbiome taxonomic profiles. While our index is explicitly trained to distinguish between healthy and diseased gut microbiomes, it also provides a practical approach to approximating pre-diseased states. This is achieved by interpolating between the healthy and diseased states, allowing GMWI2 to reveal variations across the gut microbiome health spectrum. Specifically, assuming sufficient prediction quality of our model, an individual's GMWI2 score will decrease as they transition from healthy to pre-diseased to diseased states, or increase if transitioning in the reverse direction. Moreover, GMWI2 provides a pragmatic alternative to the resource-intensive collection of longitudinal gut microbiome datasets needed to precisely track the steady transition from healthy to diseased. Current efforts in this area are very limited in scale and costly.

In all, GMWI2 is not intended for confirming specific disease diagnoses but rather serves as an early warning system, akin to a "canary in a coal mine". It is designed to detect potentially adverse shifts in overall gut health before specific, diagnosable symptoms occur. Such detection could inform dietary or lifestyle modifications to prevent mild issues from escalating into severe health conditions, or prompt further diagnostic tests. Unlike existing disease-specific indices, our index spans multiple diseases, thereby emphasizing a pan-disease (or alternatively, a generally healthy) gut microbiome signature. This broad applicability could be particularly useful in clinical scenarios such as selecting FMT donors, where gut health could be taken as a reflection of overall health. In conditions like rheumatoid arthritis and other autoimmune inflammatory disorders, GMWI2 could guide decisions on tapering or discontinuing therapy, or assessing the possibility of disease flares. In this sense, GMWI2 may potentially usher in a 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 models has the potential to open exciting possibilities for healthy aging[43] and preventative health screening and wellness programs[44,45], driven by insights from the gut microbiome.

Methods

Multi-study pooling of human stool metagenomes

We conducted a comprehensive literature search using targeted keywords such as "gut microbiome", "stool metagenome", and "whole-genome shotgun" in PubMed and Google Scholar. The search was performed up until January 2022 to identify published studies that included publicly available shotgun metagenomic data of human stool samples, along with corresponding subject meta-data. In cases where multiple samples were collected from individuals across different time points, we included only the first or baseline sample from that study subject. Studies involving dietary or medication interventions were not included in the pooled dataset for GMWI2 training. Studies with fewer than 40 samples were also excluded from our analysis, considering the potential limitations in the robustness and reliability of microbiome data from such pilot-scale microbiome studies. The raw sequence files (in .sra or .fastq format) were retrieved from the NCBI Sequence Read Archive and European Nucleotide Archive databases for further analysis.

Stool metagenome sample exclusion criteria

To minimize potential bias and preserve data integrity, we applied stringent criteria to the stool shotgun metagenome samples for inclusion in our study. Specifically, we excluded samples sequenced using non-Illumina platforms, such as 454 GS FLX Titanium, Ion Torrent PGM, Ion Torrent Proton, and BGISEQ-500, to ensure consistency in sequencing technology. In terms of data quality, we excluded samples with low read counts (below 1 million reads) prior to quality control filtration. Additionally, our analysis did not include samples from studies with a primary focus on the virome or those where stool samples underwent virus-like particle purification.

Furthering our strict sample control standards, we also excluded disease control samples that were not specifically tied to a clinical diagnosis in the originating study. Individuals who were not clinically diagnosed with a specific disease but exhibited certain anomalous conditions were also excluded. These conditions comprised: (i) a Body Mass Index (BMI) suggestive of being underweight (BMI < 18.5), overweight (BMI ≥ 25 and <30), or obese (BMI ≥ 30) were not classified as a non-healthy phenotype; (ii) declared heavy drug use (including alcohol and recreational drugs); (iii) age exceeding 100 years; and (iv) individuals initially healthy at baseline, but later reported to develop a disease condition during a longitudinal study. Additionally, samples from newborn, infant, and child gut microbiome studies were excluded since the primary focus was on adult human gut microbiomes. Lastly, we excluded non-healthy individuals with early-stage diseases (e.g., impaired glucose tolerance, hypertension, colorectal adenoma), rare or genetically-linked disorders (e.g., Behcet's disease, schizophrenia), and non-colon cancers (including pancreatic, non-small cell lung, and breast cancer). These exclusions were applied to ensure a uniform and representative dataset for training GMWI2.

Quality control of sequenced reads

Potential human contamination was filtered out by removing reads that aligned to the human genome (reference genome GRCh38/hg38) using Bowtie2[46] v2.4.4 with default parameters. Along with Illumina universal adapter sequences, probable adapter sequences were identified by extracting overrepresented sequences from each metagenome sample using FastQC[47] v0.11.8. Adapter sequence clipping and quality filtration were performed using Trimmomatic[48] v0.39. Specifically, Trimmomatic's "ILLUMINACLIP" step was used, using a maximum seed mismatch count of 2, palindrome clip threshold of 30, simple clip threshold of 10, and minimum adapter length of 2 bp. Additionally, leading and trailing low-quality bases (Phred quality score < 3) of each read were removed, and trimmed reads shorter than 60 bp in nucleotide length were discarded.

Taxonomic profiling

After performing quality filtration on all raw metagenomes, taxonomic profiling was carried out using the MetaPhlAn3[16] v3.0.13 phylogenetic clade identification pipeline using default parameters. Briefly, MetaPhlAn3 classifies metagenomic reads to taxonomies based on a database (mpa_v30_CHOCOPhlAn_201901) of clade-specific marker genes. Once taxonomic features (or clades) of unknown/unclassified identity were removed, the remaining clades that could be detected in at least one metagenome sample in the pooled dataset were considered for further analysis.

After taxonomic profiling, the following metagenomes were discarded from our analysis: (i) samples composed of >90% unmapped reads; (ii) samples with a relatively high proportion (>25%) of unknown taxa; and (iii) samples lacking sufficient taxonomic diversity (<100 identified taxa). These samples were removed to maintain the quality and reliability of the training data. Finally, after applying all exclusion criteria, studies with fewer than 20 remaining samples were removed.

Generating presence/absence taxonomic profiles

To mitigate concerns related to the compositional nature of microbiome data[49], batch effects, and to simplify the interpretation of the GMWI2 classification model, we transformed the taxa relative abundances from MetaPhlAn3 into a binary presence/absence profile for each metagenome sample. Specifically, a taxon was deemed "present" in a given sample if its relative abundance in a sample was equal to or greater than 0.00001 (or 0.001%), and considered absent otherwise. Consequently, each sample was represented as a binary vector.

PCA and PERMANOVA analysis on taxonomic profiles

Principal component analysis (PCA) was conducted on the presence/absence taxonomic profiles using the "prcomp" function in R. Additionally, Bray-Curtis distance matrices were generated based on the relative abundances of microbial taxa (ranging from phylum to species) in the stool metagenomes. This was done using the "vegan" package v2.6.4 in R. We then carried out permutational multivariate analysis of variance (PERMANOVA) on the distance matrix using the "adonis2" function. To evaluate the influence of the subjects' health status on the total variance in gut microbial community composition, we calculated the *P*-value for the test statistic (pseudo-F) based on 999 permutations.

Estimating disease likelihood using Lasso-penalized logistic regression

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 8069 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 optimal 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.

Let $\boldsymbol{x}_i$ be a binary vector encoding the presence or absence of n taxa in the ith labeled sample:

$$\boldsymbol{x}_i = \left[x_i^1, x_i^2, \cdots, x_i^n\right] \tag{1}$$

where x_i^j is 1 if taxa j is present in sample i and 0 otherwise. Additionally, $n = 3200$ is the number of taxonomic features (or clades) in the i^{th} sample (a total of 3200 taxonomic features were observed at least once in the pooled metagenome dataset).

Let y_i represent the health status (1 for healthy, 0 for non-healthy) of sample $\boldsymbol{i}$. The subsequent log-loss optimization objective function is solved using L1 regularization and class proportion weights as follows:

$$\theta^* = \underset{\theta\in\mathbb{R}^n}{\operatorname{argmin}}\, C\sum_{i=1}^{m}\alpha(y_i)\left[(-y_i\log(h_\theta(\boldsymbol{x}_i)) - (1-y_i)\log(1-h_\theta(\boldsymbol{x}_i))\right] + \parallel\theta\parallel_1 \tag{2}$$

where θ^* refers to the learned coefficient vector, C is the aforementioned inverse regularization strength parameter, $m = 8069$ represents the total number of samples in the pooled metagenome dataset, α is the class proportion weight term, and $h_\theta(\boldsymbol{x}_i)$ is the hypothesis function:

$$h_\theta(\boldsymbol{x}_i) = P(y_i = 1|\boldsymbol{x}_i,\theta) = \sigma\left(\theta^T\boldsymbol{x}_i\right) = \frac{1}{1+e^{-\theta^T\boldsymbol{x}_i}} \tag{3}$$

where σ is the sigmoid function. The class proportion term α accounts for the relatively unbalanced class proportions in the pooled dataset:

$$\alpha(y_i) = \frac{m}{2\sum_{j=1}^{m}\left[y_i = y_j\right]} \tag{4}$$

Using GMWI2 as a stool metagenome-based health status classifier

We calculated GMWI2 scores for all 8069 stool metagenomes in the pooled dataset, as well as samples from the four gut microbiome case studies. The taxonomic profile of a metagenome was represented as a vector $\boldsymbol{x}_{\text{test}}$, with binary values that encoded the presence or absence of microbial taxa. The computation employed the predicted log odds (logit) using the previously learned coefficient vector θ^*:

$$\text{GMWI2}(\boldsymbol{x}_{\text{test}}) = \left(\theta^*\right)^T\boldsymbol{x}_{\text{test}} \tag{5}$$

For classification purposes, a predetermined magnitude cutoff parameter c was utilized ($c = 0$ in case of having no cutoff or defer option). Finally, GMWI2 was computed on a metagenome $\boldsymbol{x}_{test}$ while applying the following criteria:

$$\text{classify}(\boldsymbol{x}_{\text{test}}) = \begin{cases} \text{non-healthy} & \text{GMWI2}(\boldsymbol{x}_{\text{test}}) < -c \\ \text{defer} & -c \le \text{GMWI2}(\boldsymbol{x}_{\text{test}}) \le c \\ \text{healthy} & \text{GMWI2}(\boldsymbol{x}_{\text{test}}) > c \end{cases} \tag{6}$$

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."

Evaluation of classification performance

Balanced accuracy, defined as the average of the proportions of correctly classified healthy and non-healthy samples, was used to evaluate the performance of the GMWI2 classification model. This was done across different cutoff parameters (c) using multiple validation techniques: training on the entire dataset and then testing on the same training set, 10-fold cross-validation (10-fold CV), and leave-one-out cross-validation (LOOCV). In order to account for variability in 10-fold cross-validation, the process was repeated 10 times with shuffled fold partitions, and the results were averaged across all runs. Additionally, inter-study validation (ISV) was conducted, in which a single study was held out each time, the model was trained on the remaining studies, and testing was performed on the samples of the single-held-out study. ISV allows for an assessment of classification performance across different studies.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Information regarding the stool metagenome samples (and their corresponding studies) used to train the GMWI2 classifier is available in Supplementary Data 2. Raw metagenomic reads are available using the sequencing data accession IDs.

Code availability

A command-line tool for computing the GMWI2 score of a stool metagenome from its corresponding raw .fastq sequence file can be installed via Anaconda (https://anaconda.org/bioconda/GMWI2). The source code for the tool, processed datasets (including the 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.

References

1. Schirmer, M. et al. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* **167**, 1125–1136.e8 (2016).
2. Halfvarson, J. et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* **2**, 1–7 (2017).
3. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
4. Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
5. Mars, R. A. T. et al. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell* **183**, 1137–1140 (2020).
6. Mou, Y. et al. Gut Microbiota interact with the brain through systemic chronic inflammation: Implications on neuroinflammation, neurodegeneration, and aging. *Front. Immunol.* **13**, 796288 (2022).
7. iMSMS Consortium. Gut microbiome of multiple sclerosis patients and paired household healthy controls reveal associations with disease risk and course. *Cell* **185**, 3467–3486.e16 (2022).
8. Ferreiro, A. L. et al. Gut microbiome composition may be an indicator of preclinical Alzheimer's disease. *Sci. Transl. Med.* **15**, eabo2984 (2023).
9. Morton, J. T. et al. Multi-level analysis of the gut-brain axis shows autism spectrum disorder-associated molecular and microbial profiles. *Nat. Neurosci.* **26**, 1208–1217 (2023).
10. Gupta, V. K. et al. A predictive index for health status using species-level gut microbiome profiling. *Nat. Commun.* **11**, 4635 (2020).
11. Chang, D., Gupta, V. K., Hur, B., Cunningham, K. Y. & Sung, J. GMWI-webtool: a user-friendly browser application for assessing health through metagenomic gut microbiome profiling. *Bioinformatics* **39**, btad061 (2023).
12. Gacesa, R. et al. Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**, 732–739 (2022).
13. Xu, Q. et al. Metagenomic and metabolomic remodeling in nonagenarians and centenarians and its association with genetic and socioeconomic factors. *Nat. Aging* **2**, 438–452 (2022).
14. Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C. & Knight, R. Human-associated microbial signatures: examining their predictive value. *Cell Host Microbe* **10**, 292–296 (2011).
15. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* **12**, e1004977 (2016).
16. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, e65088 (2021).
17. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
18. Leek, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
19. Sung, J., Wang, Y., Chandrasekaran, S., Witten, D. M. & Price, N. D. Molecular signatures from omics data: from chaos to consensus. *Biotechnol. J.* **7**, 946–957 (2012).
20. Pattee, J. & Pan, W. Penalized regression and model selection methods for polygenic scores on summary statistics. *PLoS Comput. Biol.* **16**, e1008271 (2020).
21. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
22. Ma, S. et al. Measuring the effect of inter-study variability on estimating prediction error. *PLoS ONE* **9**, e110840 (2014).
23. Goll, R. et al. Effects of fecal microbiota transplantation in subjects with irritable bowel syndrome are mirrored by changes in gut microbiome. *Gut Microbes* **12**, 1794263 (2020).
24. Woodworth, M. H., Carpentieri, C., Sitchenko, K. L. & Kraft, C. S. Challenges in fecal donor selection and screening for fecal microbiota transplantation: a review. *Gut Microbes* **8**, 225–237 (2017).
25. Duvallet, C. et al. Framework for rational donor selection in fecal microbiota transplant clinical trials. *PLoS ONE* **14**, e0222881 (2019).
26. Tanes, C. et al. Role of dietary fiber in the recovery of the human gut microbiome and its metabolome. *Cell Host Microbe* **29**, 394–407.e5 (2021).
27. Gibson, G. R. & Roberfroid, M. B. Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. *J. Nutr.* **125**, 1401–1412 (1995).
28. Anderson, J. W. et al. Health benefits of dietary fiber. *Nutr. Rev.* **67**, 188–205 (2009).
29. Venter, C. et al. Role of dietary fiber in promoting immune health—An EAACI position paper. *Allergy* **77**, 3185–3198 (2022).
30. Palleja, A. et al. Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nat. Microbiol.* **3**, 1255–1265 (2018).
31. Lee, D. H. et al. Evaluating the prebiotic effect of oligosaccharides on gut microbiome wellness using in vitro fecal fermentation. *Npj Sci. Food* **7**, 18 (2023).
32. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
33. Ananthakrishnan, A. N. et al. Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. *Cell Host Microbe* **21**, 603–610.e3 (2017).
34. Hjorth, M. F. et al. Prevotella-to-Bacteroides ratio predicts body weight and fat loss success on 24-week diets varying in macronutrient composition and dietary fiber: results from a post-hoc analysis. *Int. J. Obes.* **43**, 149–157 (2019).
35. Gupta, V. K. et al. Gut microbial determinants of clinically important improvement in patients with rheumatoid arthritis. *Genome Med.* **13**, 149 (2021).
36. Wilmanski, T. et al. Gut microbiome pattern reflects healthy ageing and predicts survival in humans. *Nat. Metab.* **3**, 274–286 (2021).
37. Jian, C. et al. Gut microbiota predicts body fat change following a low-energy diet: a PREVIEW intervention study. *Genome Med.* **14**, 54 (2022).
38. Sung, J. et al. Multi-study integration of brain cancer transcriptomes reveals organ-level molecular signatures. *PLoS Comput. Biol.* **9**, e1003148 (2013).
39. Parsana, P., Amend, S. R., Hernandez, J., Pienta, K. J. & Battle, A. Identifying global expression patterns and key regulators in epithelial to mesenchymal transition through multi-study integration. *BMC Cancer* **17**, 447 (2017).
40. Xu, J. et al. Algorithmic fairness in computational medicine. *EBioMedicine* **84**, 104250 (2022).
41. Herbei, R. & Wegkamp, M. H. Classification with Reject Option. *Can. J. Stat.* **34**, 709–721 (2006).
42. Hanczar, B. & Dougherty, E. R. Classification with reject option in gene expression data. *Bioinformatics* **24**, 1889–1895 (2008).
43. Ghosh, T. S., Shanahan, F. & O'Toole, P. W. Toward an improved definition of a healthy microbiome for healthy aging. *Nat. Aging* **2**, 1054–1069 (2022).
44. Price, N. D. et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat. Biotechnol.* **35**, 747–756 (2017).
45. Shen, X. et al. Multi-omics microsampling for the profiling of lifestyle-associated changes in health. *Nat. Biomed. Eng.* **8**, 1–19 (2024).
46. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
47. Wingett, S. W. & Andrews, S. FastQ Screen: a tool for multi-genome mapping and quality control. *F1000Res.* **7**, 1338 (2018).
48. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

49. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
50. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: a library for large linear classification. https://www.csie.ntu.edu.tw/~cjlin/papers/liblinear.pdf (2008).

Acknowledgements

We thank Professor Dan Knights for helpful feedback regarding this study. This work was supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health Award Numbers UL1TR002494 and UL1TR002377. Additional support was provided by the Minnesota Partnership for Biotechnology and Medical Genomics through the Translational Product Development Fund (to J.S.), as well as Mark E. and Mary A. Davis to Mayo Clinic Center for Individualized Medicine (to J.S.).

Author contributions

D.C., V.K.G. and J.S. developed the study idea and designed all analytical methodologies. D.C. and V.K.G. performed the computational experiments. All authors (D.C., V.K.G., B.H., S.C.-L., K.Y.C., N.H., I.L., V.L.K., L.M.T., L.V.K., E.E.L., J.M.D., H.N. and J.S.) analyzed and discussed the data. D.C., V.K.G. and J.S. wrote the manuscript, with contributions from other authors. All authors critically reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-51651-9.

Correspondence and requests for materials should be addressed to Jaeyun Sung.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

[1]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA. [2]Microbiomics Program, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA. [3]Viral Information Institute, San Diego State University, San Diego, CA, USA. [4]Bioinformatics and Computational Biology Program, University of Minnesota, Minneapolis, MN, USA. [5]Brain Korea 21 Center for Bio-Health Industry, Department of Food Science and Biotechnology, Chungbuk National University, Cheongju, South Korea. [6]Department of Biotechnology, Yonsei University, Seoul, South Korea. [7]Division of Rheumatology, Department of Medicine, Mayo Clinic, Rochester, MN, USA. [8]Department of Food Science and Nutrition, University of Minnesota, St. Paul, MN, USA. [9]Department of Pulmonary & Critical Care, Mayo Clinic, Rochester, MN, USA. [10]Department of Neurology, Yale University, New Haven, CT, USA. [11]Emeritus, Department of Surgery, Mayo Clinic, Rochester, MN, USA. [12]Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA. [13]These authors contributed equally: Daniel Chang, Vinod K. Gupta. ✉e-mail: Sung.Jaeyun@mayo.edu