Leveraging DNA methylation to create Epigenetic Biomarker Proxies that inform clinical care: A new framework for Precision Medicine

Natàlia Carreras-Gallo[2]+,Qingwen Chen[1]+, Laura Balagué-Dobón[2], Andrea Aparicio[1], Ilinca M. Giosan[2], Rita Dargham[2], Daniel Phelps[2], Tao Guo[1], Kevin M. Mendez[1], Yulu Chen[1], Athena Carangan[2], Srikar Vempaty[2], Sayf Hassouneh[2], Michael McGeachie[1],Tavis Mendez[2], Florence Comite[3], Karsten Suhre[4], Ryan Smith[2], Varun B. Dwaraka[2]*,Jessica A. Lasky-Su[1]*

[1] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[2] TruDiagnostic, Inc., Lexington, KY USA
[3] Comite Center for Precision Medicine & Health, New York, NY, United States
[4] Bioinformatics Core, Weill Cornell Medicine-Qatar, Education City, 24144 Doha, Qatar

+ Authors who contributed equally to this work

***Corresponding authors:**
Jessica A. Lasky-Su, Sc.D.
Channing Division of Network Medicine
Department of Medicine,
Brigham and Women's Hospital
181 Longwood Avenue, Boston, MA 02115
Email: rejas@channing.harvard.edu

Varun B. Dwaraka, Ph.D.
TruDiagnostic Inc.
881 Corporate Drive, Lexington, KY 40503
Email: varun.dwaraka@trudiagnostic.com

ABSTRACT

The lack of accurate, cost-effective, and clinically relevant biomarkers remains a major barrier to incorporating omic data into clinical practice. Previous studies have shown that DNA methylation algorithms have utility as surrogate measures for selected proteins and metabolites. We expand upon this work by creating DNAm surrogates, termed epigenetic biomarker proxies (EBPs), across clinical laboratories, the metabolome, and the proteome. After screening >2,500 biomarkers, we trained and tested 1,694 EBP models and assessed their incident relationship with 12 chronic diseases and mortality, followed up to 15 years. We observe broad clinical relevance: 1) there are 1,292 and 4,863 FDR significant incident and prevalent associations, respectively; 2) most of these associations are replicated when looking at the lab-based counterpart, and > 62% of the shared associations have higher odds and hazard ratios to disease outcomes than their respective observed measurements; 3) EBPs of current clinical biochemistries detect deviations from normal with high sensitivity and specificity. Longitudinal EBPs also demonstrate significant changes corresponding to the changes observed in lab-based counterparts. Using two cohorts and > 30,000 individuals, we found that EBPs validate across healthy and sick populations. While further study is needed, these findings highlight the potential of implementing EBPs in a simple, low-cost, high-yield framework that benefits clinical medicine.

Keywords: Epigenetics, DNA methylation, biomarkers, biological aging, proteomics, metabolomics

INTRODUCTION

The advent of multiomics has brought great potential to transform patient treatment into a tailored “personalized” framework. The molecular data extracted from regular, deep omic profiling of patients provides an unparalleled opportunity to provide nuanced patient feedback that may ultimately lead to a prolonged health lifespan. Multiomic profiles provide immense insight into the physiologic processes contributing to health. Genetic profiles capture baseline risk, epigenetics regulate gene expression, transcriptomics provides the readout of gene expression, proteomics informs upon immune and inflammatory profiles, metabolomics provides insight into metabolic processes, microbiomics describes underlying gut health, while exposomics captures a range of environmental and lifestyle data including diet, exercise, and sleeping patterns. Yet, to date, the practical clinical impact of multiomics has fallen short of the transformative revolution once anticipated in healthcare and has not yet provided a clear pathway toward precision medicine. A major drawback is that omic profiling does not fit into the context of the current clinical system. It is time-consuming, impractical on a large scale, requires repeated visits for sample draws, and involves prohibitive costs at an individual patient scale. These critical barriers must be resolved in order for broad-scale adoption among practitioners and the full realization of its potential. Nevertheless, as the scope of multiomics continues to expand, capitalizing on the knowledge embedded in omics provides a novel opportunity to address these barriers and develop a low-cost, reliable framework, with unprecedented opportunity to move the needle on precision medicine implementation.

Among the layers of multi-omics, epigenetic changes stand out for regulating gene expression while being influenced by environmental and lifestyle factors including diet, exercise, smoking, and alcohol consumption[1,2]. DNA methylation (DNAm) patterns have shown to provide unique dynamic insight into physiological responses to environmental stressors across the lifespan and are considered promising predictive biomarkers across a broad spectrum of health outcomes [2–5]. The DNAm-based research has led to significant advancements in predicting biological aging and disease risk through the development of robustly validated phenotypic algorithms such as biological age clocks and methylation risk scores (MRSs)[6,7]. To date, there are over 30 biological age clocks that strive to quantify aging on a molecular level and have proven to be highly predictive of health outcomes. MRSs have demonstrated high accuracy in diagnosing and predicting disease, often outperforming polygenic risk scores (PRSs)[6,7].

In recent years, DNAm data has been further used to create surrogate measures (a.k.a. DNAm scores, DNAm surrogates, EpiScores) of proteins and metabolites, enabling access to otherwise costly or inaccessible clinical metrics. DNAm algorithms have been developed to estimate selected blood proteins and metabolites, capturing shifts in inflammation, metabolic function, and environmental exposures [8–10,11,12]. Recent work has demonstrated the ability of DNAm surrogates to reflect relevant biology and serve as markers of disease, augmenting the current clinical biomarkers for risk stratification[,11,13,14]. Further, DNAm surrogates have been integrated into the development of biological aging clocks, showing improved predictive accuracy for longevity and strong associations with age-related chronic disease outcomes[3,13,14], and their clinical relevance has led to the development of diagnostic tests for Mendelian genetic disorders[15]. While the ability of DNAm surrogates to capture relevant biology has been established, a comprehensive assessment of their validity across a spectrum of molecular data and clinical outcomes remains unrealized. Conducting such an assessment using large-scale independent cohorts is crucial for establishing their clinical relevance.

Here, we address this gap by building DNAm surrogates, termed epigenetic biomarker proxies (EBPs) of ~1,600 clinical, metabolomic, and proteomic measurements for over 4000 participants in the Massachusetts General Brigham's Aging Biobank Cohort (MGB-ABC)[13]. We demonstrate that the EBPs are strongly correlated with the corresponding actual measurements, and that they often have even stronger associations with disease outcomes than their lab-based counterparts. Through analyses that include gene ontology, association testing, and longitudinal analysis, we identify sets of EBPs that have important clinical relevance, and can inform upon underlying pathobiology. Collectively, our results show the numerous advantages that EBPs can offer in a clinical context. Namely, they are able to capture biological insights with very good precision at a fraction of the cost of the composite of clinical and omic measurements. They can therefore provide comprehensive and personalized patient feedback in a framework where blood samples can be readily collected at home, even in remote areas. We envision that the use of EBPs in clinical settings represents a cost-effective and accessible pathway into a new precision medicine paradigm.

RESULTS

Study design

To develop and validate the EBPs, we used 4,418 participants from the Massachusetts General Brigham's Aging Biobank Cohort (MGB-ABC). All participants were assessed using DNA methylation profiling and subsets of them were assessed using proteomics, metabolomics, and clinical lab tests. Untargeted global plasma metabolomic profiling was performed on the Metabolon platform consisting of 1,150 metabolites across 1,148 individuals. Global proteomic data were generated using the Seer SP100 platform, based on liquid chromatography-mass spectrometry, consisting of 28,317 non-unique protein groups and 1,979 unique protein groups (denoted as 'proteins') across 1,260 individuals. Finally, 44 clinical lab tests were extracted from EMR files through the Research Patient Data Registry portal[16]. Since clinical lab tests were not conducted for all patients at the time of biosample collection, only the subset of patients with available lab values was included in the development of clinical EBPs. To assess the effectiveness of the developed EBPs, we used an independent cohort, the TruDiagnostic Biobank, which consists of 31,012 samples from 26,155 individuals tested on the EPICv1 (15,597 samples) or the EPICv2 (15,415) and who have consented for research. **Fig. 1A** describes the general characteristics and demographics of the cohorts. The additional characteristics are described in **Supplementary Table S1**.

Development of EBPs

We developed a machine learning pipeline to generate EBPs for each of 1,150 metabolites from the Metabolon platforms, 1,979 proteins from the Seer platform, and 44 clinical lab tests obtained from electronic medical records. The pipeline comprises two main steps. First, a filter selects the top 10% of CpG sites with the highest mutual information with the target, effectively reducing the feature space. Following feature selection, the dataset is split into training and testing sets in an 85:15 ratio. Using the training data, we fitted elastic-net linear regression (ENET) models, evaluating four different alpha values (0.01, 0.1, 0.5, 1). Across all alpha values, we retained the optimized model that exhibited the lowest mean squared error (MSE). In cases where the predictions of the selected ENET model were weakly correlated with observed values (Spearman correlation < 0.2) in the testing set, we applied XGBoost as an additional step. The XGBoost model was retained only if the updated predictions achieved a Spearman correlation ≥ 0.2 with the observed values. In total, 1,694 effective biochemical predictors (EBPs) were retained, comprising 689 metabolite EBPs, 963 protein EBPs, and 42 clinical EBPs (**Fig. 1B**).

Correlation between EBPs, the observed biomarkers, and prior DNAm surrogates

To evaluate the performance of the EBPs, we calculated their Spearman correlation to the observed measures in the testing dataset (**Supplementary Fig. 1A**). The metabolite EBPs had a mean correlation of 0.29, with a range between 0.20 and 0.59 (**Supplementary Table S2**). The metabolites with the highest correlations to the observed values were androstenediol monosulfate (corr = 0.59), 5alpha-androstan-3alpha,17beta-diol monosulfate (corr = 0.58), and C-glycosyltryptophan (corr = 0.57), while the bottom metabolites were cyclo(gly-pro) (corr = 0.20), glycolithocholate (corr = 0.20), and N-acetylasparagine (corr = 0.20). Splitting metabolites further into annotated super pathways (**Supplementary Fig. 1B)** resulted in Nucleotide associated metabolite EBPs showing the greatest mean correlation (mean Spearman = 0.32), while EBPs associated with Energy pathways showed the lowest (mean Spearman = 0.23). The protein EBPs had a mean correlation of 0.29, with a range between 0.20 and 0.54 (**Supplementary Table S3**). The proteins with the highest correlations were HLA class I histocompatibility antigen, alpha chain G (corr = 0.54), Semaphorin-3F (corr = 0.54), and Collectin-11 (corr = 0.52), while the bottom proteins were Extended synaptotagmin-1 (corr = 0.20), AMP deaminase 2 (corr = 0.20), and Elafin (corr = 0.20). The clinical lab EBPs had the highest overall correlations, with an average correlation of 0.41 and a range from 0.23 to 0.67 (**Supplementary Table S4**). The clinical values EBPs with the highest correlations to the observed values were glomerular filtration rate (corr = 0.67), lymphocyte count (corr = 0.64), and neutrophil count (corr = 0.64), while the bottom observed clinical values were thyroid-stimulating hormone (corr = 0.23), total bilirubin (corr = 0.23), and calcium (corr = 0.23). In total, we found 21 clinical EBPs, 71 metabolite EBPs, and 153 protein EBPs with a correlation greater than 0.4. The total number imbalance, despite the clinical EBPs having the highest overall correlation, is due to the larger number of observed proteins and metabolites, than of clinical values.

We compared the correlations between the EBPs developed here, the observed biomarker values, and prior DNAm surrogates, including 18 Protein EpiScores[11], 6 Nightingale metabolite surrogates[17], and 5 principal-component (PC) based surrogate proteins from GrimAge[3,18] (**Supplementary Fig. 2**). We found that, overall, the EBPs showed a stronger correlation with the observed proteins and metabolites (excluding the protein CXCL10 and the metabolite leucine) and were correlated with prior DNAm surrogates.

Incident and prevalent association of EBPs and Observed Biomarkers to 12 chronic diseases and mortality

We evaluated the relationship between EBPs, observed biomarkers, and 12 major diseases (cancer, COPD, asthma, cardiovascular disease - CVD, congestive heart failure, coronary artery disease, stroke, type 2 diabetes, chronic kidney disease, depression, cognitive deficit, chronic liver disease) and all-cause mortality in the MGB-ABC cohort. We assessed both incident and prevalent associations for each EBP-disease and observed value-disease pair, adjusting for sex and age, at a False Discovery Rate (FDR) significance level of 0.05 (**Table 1, Supplementary Table S5**). Given the potential for confounding by indication with prevalent diseases, we focus on the incident disease associations; however, we provide the analytic results for both prevalent and incident disease associations for reference.

There were 1,292 significant disease-EBP and 1,063 significant disease-observed biomarker associations. The EBPs captured a total of 72.3% (769/1063) of the significant associations identified with the observed biomarkers, representing 90.6%, 82.6%, and 64.8% of the associations with the clinical outcome, protein, and metabolite biomarkers, respectively. Among the remaining 27.7% (294/1063) significant disease-observed biomarker associations that were not captured with the EBPs associations, 234 were metabolite EBPs, followed by 49 protein EBPs, and 11 clinical EBPs. Among the 769 significant associations observed with both EBPs and the observed biomarkers, 62% of the EBPs had higher HRs to disease when compared with the observed biomarkers. This varied across disease and EBP type: 81%, 64%, and 56% of the clinical, protein, and metabolite EBPs, respectively, had higher HRs than their observed counterparts. This also varied by disease where EBPs had higher HRs than the observed values for 9 of the 13 outcomes, including mortality, asthma, liver, diabetes, cognition, and all cardiovascular diseases. Volcano plots comparing the statistically significant disease-EBP and disease-observed biomarkers for each disease and all-cause mortality (**Supplementary Fig. 3-16)** show consistency among the highest EBPs and observed biomarkers for each disease. **Fig. 2** shows the top 10 biomarkers for each disease looking at the observed values and the EBPs. We observed high consistency in the biomarkers, with as many as 8 out of 10 consistent between EBPs and observed values for Congestive Heart Failure. To compare the magnitude of the effects of predicted EBPs and observed values on disease, we generated forest plots displaying the top 20 EBPs with the highest and lowest HRs and ORs for each disease along with their matching observed values (**Supplementary Fig. 17-29**). We confirmed that EBPs aligned with diseases in the same direction as observed values and, often, EBPs had higher effect estimates. It is important to note that our prevalent EBP analysis might suffer from confounding by indication of disease, such as medications that manage specific health conditions. As such, some of the prevalent disease associations may be misleading. Such an example can be observed with depression (**Supplementary Fig. 13B**), where we see the most significant metabolite and DNAm EBP for prevalent disease as being serotonin. It is possible that this is not a signal of disease but a signal of treatment such as serotonin reuptake inhibitors (SSRIs).

Multimorbidity of EBP and observed biomarker association findings

As expected, we identified many EBPs with multimorbidity across the chronic diseases and mortality (**Fig. 3**), reflecting common physiologic responses of poor health. Overall, the EBPs exhibited similar multimorbidity to the respective actual biomarkers. expected in their observed counterparts due to their involvement in different biological pathways. For example, red cell distribution width, BMI, glucose, or hematocrit are known for their association with cardiovascular diseases, type 2 diabetes, and cancer, which is replicated by their corresponding EBPs. In total, 16 clinical, 13 metabolite, and 4 protein EBPs had significant incident associations with 5 or more diseases, while all-cause mortality had the greatest multimorbidity across all EBPs (**Fig. 3A**). The strong multimorbidity for several of the clinical EBPs is not surprising, as clinical labs, including red cell distribution width, BMI, glucose, and hematocrit have established associations with many chronic diseases and and are often used as biomarkers to flag deviations from expected clinical ranges. As such, the clinical EBPs reflected what would be expected in the actual biomarkers. When looking at prevalent disease chronic kidney disease and congestive heart failure had the highest number of EBPs with significant associations for 5 or more chronic diseases (**Fig. 3B**). The EBPs associated with cancer had the lowest multimorbidity with other chronic diseases for both incident and prevalent associations.

Gene ontology enrichment analysis for clinical EBPs

To evaluate the biological processes implicated in the CpGs selected for the clinical EBPs, we performed a gene ontology (GO) enrichment analysis for these models (**Supplementary Table S6**). Many of the highly weighted CpGs retained for various EBPs included gene regions regulating biochemical processes relevant to the estimated biomarkers. For example, alkaline phosphatase (ALP), systolic blood pressure, and hematocrit all had enrichment of biochemical processes and pathways related to the physiology of each biomarker. ALP GO enrichment analysis identified that negatively weighted CpG sites were enriched in cellular responses to corticosteroid/glucocorticoid stimulus and acid chemical (**Fig. 4A**). Glucocorticoids regulate ALP by activating its production[19]. Two of the genes where these CpG sites were mapped were *SMAD3* and *CD8A*, both of them are related to ALP activity and corticosteroid/glucocorticoid pathways[20–23]. For hematocrit, we found that genes from negatively weighted CpG sites were enriched in hematopoiesis, and leukocyte differentiation (**Fig. 4B**). Finally, the negatively weighted CpG sites for systolic blood pressure were enriched in positive regulation of peptide and peptide hormone secretion (**Fig. 4C**).

Association of EBPs in Disease and Supplementation

To evaluate whether EBPs reflect known associations observed in other cohorts, we examined their relationship with various conditions, lifestyle factors, and supplementation patterns. Using 31,012 samples from the TruDiagnostic cohort and self-reported patient questionnaires collected at sample registration, we assessed 150 predefined associations (**Supplementary Table S7**). These associations spanned 10 lifestyle factors (e.g., tobacco use, alcohol consumption, and primary dietary patterns), 37 diseases (e.g., hypertension, cancer, and gout), and 20 types of supplementation (e.g., vitamins, Omega-3, and dietary supplements). Given the significant sex differences observed in many EBP levels, all analyses were adjusted for sex, with additional stratified analyses conducted for males and females independently (**Supplementary Table S8**).

Gout is a painful form of arthritis caused by uric acid accumulation, a byproduct of purine metabolism. To explore the role of DNAm in gout, we generated violin plots and analyzed DNAm EBPs for metabolites in the uric acid pathway. As shown in **Fig. 5A-D**, sex-specific associations revealed significant increases in DNAm xanthine, urate, and allantoin in males with gout compared to those without (p = 5.8e-04, p = 3.0e-05, and p = 1.0e-19, respectively). In females, urate (p = 3.1e-05) and allantoin (p = 7.7e-07) showed significant increases, while DNAm xanthine did not (p = 0.29). These findings suggest that DNAm EBPs of uric acid metabolites may serve as useful markers for diagnosing and managing gout.

Phenylacetylglutamine, a metabolite linked to gut microbial metabolism, and vanillactate have been associated with cardiovascular disease risk, including hypertension. Consistent with prior studies, we observed significant increases in DNAm phenylacetylglutamine (p = 4.1e-128) and vanillactate (p = 2.8e-175) in patients with high blood pressure (**Fig. 5E-H**).

NAD+ supplementation, commonly used for its role in cellular metabolism and aging, was associated with specific DNAm changes. Significant increases were observed in DNAm 1-methylnicotinamide (p = 2.2e-07) and N1-methyl-2-pyridone-5-carboxamide (p = 9.2e-09) but not in Nicotinamide (p = 0.24) (**Fig. 5I-L**). Further analysis focused on NAD+ precursors, such as Nicotinamide Mononucleotide (NMN) and Nicotinamide Riboside (NR). While DNAm 1-methylnicotinamide increased with NMN supplementation (p = 0.01), significant increases were observed in DNAm 1-methylnicotinamide and N1-methyl-2-pyridone-5-carboxamide with NR supplementation (p = 1.4e-11 and p = 1.6e-12, respectively) (**Supplementary Fig. 30A-F**).

Omega-3 fatty acids, widely consumed as dietary supplements, were associated with significant increases in 3 EBPs for omega-3 fatty acids: Docosahexaenoic Acid (DHA) (p = 1.7e-10), Docosapentaenoic Acid (DPA) (p = 9.8e-08), and Eicosapentaenoic Acid (EPA) (p = 1.6e-20) in individuals reporting Omega-3 supplementation (**Supplementary Fig. 30G-I**).

Accuracy assessment of clinical EBP to detect deviations from recommended clinical ranges

To evaluate the potential usefulness of EBPs as a clinical tool, we evaluated the performance of EBPs in detecting warning signs in biomarkers levels typically tested during health screenings. To better mimic a clinical setting, we grouped the 42 clinical biomarkers into 7 panels: Cardiovascular Metrics, Complete Blood Count (CBC), Differential WBC, Glycemic Control, Inflammatory, Lipid, and Metabolic Panels (**Supplementary Table S9**). For all the patients in the test set of the MGB-ABC cohort, we first compared the observed measurements of 42 clinical biomarkers to a general clinical reference, and classified them as either *within range* or *outside of range.* Then, we compared the EBPs corresponding to each of the clinical values to the same reference (**Supplementary Fig. 31**); **Fig. 6A** illustrates this comparison for the biomarkers in the Metabolic Panel. Finally, we calculated the frequency with which an EBP correctly identified the within range or outside of range observed values, as well as the frequency with which they didn't (**Supplementary Fig. 32**). **Fig. 6B** illustrates this classification for the Creatinine EBP. Across all clinical values, we observed an overall accuracy, specificity, and sensitivity of 90%, 94%, and 76%, respectively (**Fig. 6C, Supplementary Fig. 33**). The accuracy among panels ranged from 97% (Differential WBC Panel) to 71% (Glycemic Control Panel), while the specificity ranged from 98% (Differential WBC Panel) to 21% (Glycemic Control Panel), and the sensitivity ranged from 94% (Cardiovascular Metrics Panel) to 49% (Lipid Panel). Individually, the EBPs that performed the best were Creatinine, Lymphocyte count, and Blood Urea Nitrogen, and we found a very high correlation between the proportion of out of range values in the test set, and the sensitivity and specificity (Pearson 0.88 and -0.97, respectively, **Supplementary Table S9**). In the cases where the observed biomarkers didn't have any out of range measurements, we achieved almost perfect specificity and accuracy. These results underscore the potential effectiveness of EBPs in categorizing patients within a clinical context.

Longitudinal analysis or EBPs versus Matched Observed Biomarkers

Finally, we assessed the effectiveness of EBPs in consistently following the levels of clinical biomarkers across different timepoints in a subset of 9 individuals in the TruDiagnostic cohort for which measurements of 33 clinical biomarkers and DNA methylation are available for at least 4 matched timepoints. To quantify the EBPs' performance in tracking longitudinal observed values, we first mean centered and normalized each biomarker's observed and EPS measurements, and then we calculated the difference between the EBPs and their corresponding observed value. **Fig. 6D** illustrates this difference for Biomarkers in the Metabolic Panel (the vertical axis in the panel displays the whole range of the normalized measurements). While we observed that the differences are relatively small, here the main interest lies in the variability of the trajectories; a small variability, or a stable trajectory, indicates good longitudinal tracking of the observed values by the EBPs. To quantify the trajectory variability, we calculated the variance of the differences between observed and EBP measurements across all timepoints for every individual and biomarker (**Fig. 6E**). We observe that biomarkers in the Metabolic, Lipid, and Complete Blood Count panels tended to have the smallest median variance, with a few exceptions in the first two. On the other hand, biomarkers in the Glycemic Control panel tended to have the highest variance. Some biomarkers with higher variance (glucose, calcium, triglycerides and magnesium) were amongst the least correlated with their observed values (**Supplementary Table S4**).

DISCUSSION

In this study, we developed EBPs for 1,694 circulating proteins, metabolites, and clinical biochemistries and evaluated their clinical utility in relationship with chronic diseases and mortality, followed up to 15 years. Our findings demonstrate that overall, EBPs capture the significant incident and prevalent disease associations that were observed in the actual biomarkers. We found that, when the association of EBPs and their observed counterparts with disease were both significant, more often than not, the EBPs have stronger associations than the observed biomarkers themselves. Additionally, the EBPs of clinical biochemistries detect deviations from standard clinical ranges with high accuracy and provide consistent longitudinal tracking of the observed clinical values, which was validated in an external cohort of over 30,000 individuals. To date This work highlights the potential usefulness of EBPs as a broad-scale screening tool to assess overall health. As a whole, EBPs not only provide an initial screen of standard clinical biochemistries and expand the overall scope to capture metabolic and inflammatory profiles (e.g. metabolite, protein) that provide additional insight into physiology. Additional work is needed to expand, refine and evaluate the performance of EBPs. EBPs offer a new approach to improve clinical outcomes and provide a low-cost, scalable solution for disease risk prediction and patient management.

We extended upon prior work to include a greater number and range of DNAm surrogates, increasing to >1,600 EBPs that capture clinical, proteins, and metabolites. We show that when both the observed biomarkers and EBPs were significant, in more than 62% of the associations the EBPs had higher effect size estimates with disease incidence relative to their measured counterparts. Previous work, ranging from studies of a singular protein up to 109 protein EpiScores, also observed stronger disease associations compared with the actual measured proteins and capture relevant inflammatory biology for the respective protein surrogates that were created [10,11,17,24–26]. While these increases seem counterintuitive, it may reflect differences in underlying biological signals captured through the CpGs that comprise EBPs compared with the actual measures themselves. Another study creating 64 metabolite surrogates found that these improved the prediction of mortality, but did not evaluate their association across disease outcomes[17]. By identifying biologically-relevant clinical, protein, and metabolite EBPs associated with a range of clinical outcomes, our results scale these findings by a magnitude and demonstrate that the same principles apply across a broader range of physiological processes and disease outcomes. As such, this work provides a broad-based proof of concept that protein, metabolite, and clinical EBPs capture biologically-relevant disease signals.

Additionally, we assessed the usefulness of EBPs in a clinical setting by measuring their ability to detect important deviations from clinical reference ranges and to track longitudinally with changes in clinical biomarkers. EBPs achieved an accuracy of 90% in detecting deviations from clinical reference ranges, demonstrating their potential efficiency as a convenient clinical screening tool across multiple biological processes at a fraction of the price. Notably, the EBPs for Creatinine, and Blood Urea Nitrogen performed particularly well, suggesting the possibility of using EBPs to screen patients for kidney disease, which goes largely undetected in the present [27]. We observed a very high specificity-sensitivity tradeoff, and a high correlation of these two with the proportion of out of range observed values in the cohort, suggesting the need for further validation. The majority of EBPs also showed a consistently good longitudinal tracking of the observed biomarkers, underscoring their suitability for patient supervision and followup screenings. We also showed that EBPs reflect relevant physiology for various health metrics (e.g., gout, high blood pressure) and lifestyle factors (e.g., supplement use). Furthermore, for the specific cases we selected, we observed biologically-relevant changes in EBPs with disease progression and supplement use: 1) Uric acid pathway EBP metabolites changed with gout development; 2) EBPs effectively captured alterations in NAD+ metabolites following supplementation with NAD+, NMN, and NR; 3) omega-3 fatty acid- related EBPs also increased after omega-3 supplementation. Taken together, these findings highlight the potential of EBPs for monitoring health status and guiding personalized interventions.

It is important to highlight that current clinical biomarkers provide direct measures of physiology with specific clinical implications that EBPs may not have. Thus, EBPs should be considered complementary tools rather than replacements for traditional biomarkers. Nonetheless, EBPs have an enormous potential as inexpensive screening tools to simultaneously assess multiple health outcomes and identify individuals who may benefit from further diagnostic testing. They can also serve as risk stratification and/or diagnostic/prognostic biomarkers, particularly in situations where current biomarkers are limited or invasive. Further refinement and expansion of EBPs may capture molecular patient profiles that may offer guidance for clinicians as a precision medicine tool - where none currently exists - as an approach to guide overall patient care. EBPs may also be used as a complement to other risk scores. While methylation risk scores can provide important information about overall disease risk, EBPs can often supplement this information to provide actionable recommendations for heterogeneous diseases. For example, a cardiovascular methylation risk score identifies an individual at risk, but the incorporation of lipid, inflammatory, HBA1C EBPs provides more information that may help guide clinical treatment [28].

This study also identified the relevance in a large range of protein and metabolite EBPs that may capture physiology that is relevant to overall health. A large number of protein and metabolite EBPs exhibited strong correlations with observed values and also had high multimorbidity across multiple disease outcomes, suggesting that these EBPs may be important for overall monitoring of health status. For example, the glucuronate and pseudouridine EBP metabolites were highly multimorbid and associated with five outcomes: CVD, congestive heart failure, cognitive deficit, chronic kidney disease, and all-cause mortality. These metabolites have been described in the literature as being associated with longevity and healthspan[29–31] and may be useful as markers to monitor overall well-being. The proteins, large ribosomal subunit protein and growth differentiation factor-15 were also multimorbid. Prior research has highlighted their associations with multiple diseases and identified these proteins as potential systemic biomarkers [32–34], which also may be useful for monitoring health status. These findings suggest that select EBPs, such as these, could have immediate applications to enhance patient care by complementing existing clinical labs to monitor overall health and well-being.

There are multiple bottlenecks that must be addressed in order to bring precision medicine to the forefront of clinical care, making many approaches for fast implementation challenging. First, the infrastructure for generating large-scale multi-omic data requires highly specialized facilities and personnel to generate, manage, and interpret the large amount of data produced. Therefore, the availability of this machinery is extremely limited and the associated costs are prohibitive for the average patient. Second, the integration of multi-omics to the existing clinical data infrastructure could represent an added level of complexity to a healthcare system already difficult to navigate. Third, omics-based precision medicine requires multiple and regular laboratory visits which can be impractical or impossible for some patients due to personal, social, geographical, or medical barriers. The implementation of EBPs as a screening tool represents a lower-cost, home-based alternative to overcome these barriers, providing insights into biological pathways associated with diseases and risk factors. Initial health assessment through EBPs offers the possibility of early detection of warning signals and informed targeted specialized testing, leading to timely interventions. EBPs are simple, portable, and feasible to implement, requiring only a single biosample, which reduces logistical challenges and aligns seamlessly with the existing medical framework as an LOINC-coded test similar to point-of-care (POC) diagnostics. This makes EBPs especially promising for addressing challenges in prognostics in underserved and socioeconomically disadvantaged regions, where access to regular healthcare services remains limited. Initial health assessment through EBPs offers the possibility of early detection of warning signals and informed targeted specialized testing, leading to timely interventions.

There are strengths and limitations with the current study that should be considered. The major advance is the 10 fold expansion of DNA methylation surrogates (EBPs) and the demonstration that EBPs work broadly, across multiple physiological processes that include standard clinical labs as well as a broad range of proteins and metabolites that span multiple inflammatory and metabolic processes. We demonstrate that EBPs track with longitudinal changes and have meaningful associations across multiple chronic diseases and mortality. Despite this, there are several limitations that should be considered. First, the protein EBPs and some of the metabolite EBPs were generated on global profiling platforms and therefore are not quantitative. As such, quantitative cut-offs needed to establish clinical ranges were not incorporated into this study. Future efforts expanding upon promising protein and metabolite targeted studies could follow-up on this work. Second, we observed overfitting in the elastic net and XGBoost models, likely due to the limit of sample size. However, even with this overfitting, the validation of these models against observed data highlights their robustness as viable surrogates, as they associate with disease and mortality outcomes in a manner consistent with expectations. Future efforts incorporating additional data from larger biobank cohorts could help address this limitation and further refine these models. While our study based on the MGB-ABC cohort is the first to generate and assess EBPs on the scale that was performed here, additional validation and assessments for generalizeability across multiple populations is essential to determine the universality of these findings. Further work to expand EBPs, particularly to tests that are more difficult and improve the correlation.

This study demonstrates that EBPs can capture meaningful biological signals associated with chronic diseases, providing a promising avenue for improving disease prediction, personalized medicine, and patient care. While further validation and refinement are needed, the low cost and ease of collection of EBPs make them a compelling tool for clinical applications, particularly in situations where current biomarkers are limited or invasive. Ultimately, EBPs could complement existing diagnostic strategies and enable more efficient, widespread screening for a variety of diseases, thereby reducing healthcare costs and improving patient outcomes.

MATERIALS AND METHODS

Discovery Cohorts

Massachusetts General Brigham Aging Biobank Cohort (MGB-ABC)

The Massachusetts General Brigham (MGB) Biobank is a large biorepository that provides access to research data and approximately 130,000 high-quality banked samples (plasma, serum, and DNA) from >100,000 consented patients enrolled in the MGB system. These patients can be linked to corresponding Electronic Medical Record (EMR) data, dating from the start of their medical history within the MGB network, in addition to survey data on lifestyle, environment, and family history. The patients are collected through routine healthcare visits and followed over the history of their lives.

The MGB-Aging Biobank Cohort (MGB-ABC) consists of 4,418 randomly selected participants from the MGB Biobank aimed at establishing a proportionately representative aging biobank population. Participants were selected based on even weighting in terms of age, sex and BMI, representative of the MGB Biobank. Comprehensive EMR, metabolomic profiling, proteomic profiling, and epigenetics are available for select individuals in MGB-ABC.

Blood samples were used for serum, plasma, and DNA/genomic research, and were collected either as part of clinical care or through research draws at Brigham and Women's Hospital (BWH) or Massachusetts General Hospital (MGH). Each blood draw typically involved collecting 30–50 ml of blood, and was linked to the corresponding clinical data from the Electronic Medical Record (EMR). The Biobank team also gathered additional health-related information during the blood draw process.

Questionnaires for the study were either administered electronically or in written form, and included questions related to family history, lifestyle, and environment. Participants spent approximately 10–15 minutes completing the surveys and strict measures were taken to ensure confidentiality and information security. Survey data was encrypted and participants' identities were protected by not requesting identifiable personal information.

The Phenotype Discovery Center (PDC) of MGB integrates various data sources, including Research Patient Data Registry (RPDR), health information surveys, and genotype results, into the Biobank Portal. This portal combines specimen data with EMR data, creating a comprehensive SQL Server database with a user-friendly web-based application[35]. Researchers can conduct queries, visualize longitudinal data with timestamps, employ established algorithms to define phenotypes, utilize automated natural language processing (NLP) tools for analyzing EMR data using the Informatics for Integrating Biology and the Bedside (i2b2) toolkit [36]. They can also request samples from cases and controls. Data in the Biobank Portal database includes narrative data from clinic notes, text reports (cardiology, pathology, radiology, operative, discharge summaries), codified data (e.g., demographics, diagnoses, procedures, labs and medications), and patient-reported data from the health information survey on exposures and family history. Validated phenotypes are available in the Biobank Portal user interface for genotyped Biobank participants, and other relevant measures, such as lung function, were extracted using a self-developed algorithm incorporating NLP.

TruDiagnostic Biobank Cohort

The TruDiagnostic Biobank cohort includes 31,012 samples from 26,155 participants, recruited primarily from the United States between October 2020 and September 2024, who completed the commercial TruDiagnostic TruAge test and had their DNA methylation data generated. Compared to participants from the MGB cohort, these participants were in better health. The majority of the samples were collected under a healthcare provider's recommendation or guidance while less than 5% were in a direct-to-consumer setting, possibly introducing a self-selection bias as most participants were likely to seek preventative medicine and have fewer comorbidities than normal patient populations. Participants were also asked to complete a survey about personal information, medical history, social history, lifestyle, and family history. The study involving human participants was reviewed and approved by the IRCM Institutional Review Board (IRB) and all participants provided written informed consent to take part in the study.

Profiling

Methylation profiling

The DNA methylation data for this analysis was previously described[13]. Briefly, whole blood samples were processed using the Illumina Infinium® MethylationEPIC 850K BeadChip, which offers extensive genome-wide coverage across over 850,000 methylation sites, encompassing key regulatory and functional genomic regions. Sample preparation and bisulfite conversion were conducted at TruDiagnostic Inc. under standardized conditions. Quality control and preprocessing followed established pipelines (*minfi* and *ENmix*), with stringent criteria applied to retain 4,418 high-quality samples and 863,430 reliable probes. Data normalization employed Funnorm and RCP methods to minimize inter-sample variability [37].

In the TruDiagnostic cohort, the same process was followed for processing the samples, but half of them were analyzed using the Infinium® MethylationEPIC v2 BeadChip instead of the Infinium® MethylationEPIC 850K BeadChip. To estimate the EBPs in EPICv2 samples, we imputed the CpGs that were missing from the EPICv1 array by using an age-averaged DNA methylation level based on windows of 5 years.

Metabolomic Profiling

Untargeted global plasma metabolomics data were previously generated as previously described[13]. Briefly, profiling was conducted by Metabolon Inc. using established quality control procedures, with coefficients of variation assessed through blinded QC samples to control batch effects. Metabolite extraction and profiling followed protocols as previously described, utilizing four distinct LC-MS methods tailored to detect complementary metabolite classes: (1) amines and polar metabolites in positive ion mode, (2) central and polar metabolites in negative ion mode, (3) polar and non-polar lipids, and (4) free fatty acids, bile acids, and intermediate polarity metabolites [38,39].

Metabolite identification was facilitated through an automated comparison of ion features against a reference library of ~8,000 standards, supported by manual curation for quality assurance. Quantification involved area-under-the-curve measurements normalized per run-day to address instrument tuning variability, setting median values to 1.0. Metabolite identification adhered to stringent matching criteria, with new spectral entries created for unnamed compounds, identified through consistent recurrence patterns.

Data quality control and processing applied stringent thresholds, excluding features with signal-to-noise ratios <10 or >10% missing values, with remaining gaps imputed at half the minimum peak intensity. Features with a CV >25% in pooled samples were excluded for technical consistency. The final dataset comprised 1,150 metabolites from 1,148 samples, log-transformed for normalization, and pareto scaled to harmonize measurement scales across the metabolome.

Proteomic Profiling

The proteomic data used in this study were largely the same as described previously[13]. In summary, relative protein levels were quantified from 2,000 samples (including 1,600 from the MGB-ABC cohort and 400 process controls) using the Proteograph Product Suite (Seer, Inc.) in conjunction with LC-MS. Proteins were captured using five proprietary nanoparticles on the Seer SP100 proteograph, enabling selective binding based on physicochemical properties. Captured proteins were subsequently digested with trypsin, and relative quantification was performed using the data-independent acquisition (DIA) method integrated within the Protograph Analysis Software (PAS). This produced estimations of a total of 28,490 peptides across blood samples (average 15,239), and 10,265 (average 4,281) across the controls.

The proteomic normalization procedures employed in this study were consistent with those previously detailed in the Tarkin study [40]. Briefly, the Massachusetts General Brigham (MGB) Biobank provided joint phenotype and genotype data for 1,260 samples, comprising 662 females and 598 males, aged 23–99 years (median age 70, mean 67.2), predominantly white (1,057 participants). Imputed genotype data were filtered for biallelic variants using bcftools (v1.16) and converted to plink format, with subsequent filtering parameters set as --geno 0.1, --mac 10, --maf 0.05, and --hwe 1E-15. A final set of 1,260 samples was retained with proteomic data matched to 5,461,287 genetic variants. Principal components were computed using plink2 with the --pca option.

Proteomics processing involved the Seer Proteograph workflow for plasma sample preparation and peptide generation, followed by analysis on a Bruker timsTOF Pro 2 mass spectrometer using the dia-PASEF method. Peptide and protein intensities were derived through DIA-NN (v1.8.1) with a library-free search using UniProt (UP000005640_9606) and processed with the match between runs (MBR) option. Libraries were constructed to exclude or include common protein-altering variant (PAV) peptides, detailed in prior studies.

DIA-NN's normalized intensities (PG.Normalised) served as the protein readouts. From the total 28,317 protein groups quantified across 5 nanoparticles, 6,063 were present in at least 20% of samples. For proteins detected across multiple nanoparticle runs, the run with the highest total intensity was retained. Protein data underwent log10 transformation, residualization by age, sex, and the first ten genotype principal components from the associated genotype data from the same sample, followed by inverse-normal scaling. This normalization approach supports robust proteomic analyses, building on established high-dimensional omics workflows. The final processed dataset represented 1,979 proteins.

EBP development

Feature selection

To develop DNA methylation-based predictors for specific protein levels, metabolite concentrations, and clinical lab test values, we used the normalized DNA methylation dataset, transposing it so that CpG sites were considered features. In addition, sex was encoded as a binary feature (Gender_M for males and Gender_F for females), and was included as a penalized covariate, along with chronological age. Prior to training, we first filtered CpG site features based on mutual information to retain those most informative for each target outcome. The feature selection process began with discretizing features in the complete dataset by dividing each continuous methylation variable into 15 bins, thus converting it into discrete intervals. Mutual information scores between each CpG feature and the target outcome (e.g., specific metabolite levels) were then computed using the *mutinformation* function in the R package *infotheo*. To maintain specificity, mutual information was calculated between each feature and the target while excluding the target variable itself from the calculations. To focus on highly predictive features, we applied a 90th percentile threshold, selecting only those CpG sites with mutual information scores above this cutoff. This percentile-based filtering ensured the retention of only the most relevant features, and filtered scores were saved separately for each target variable. The resulting dataset, containing the reduced CpG features, was combined with the target outcome to create the dataset. From this dataset, samples were split at random into an 85-15 train test ratio; the same train and test datasets were used for all training methods.

Elastic Net Training

For training, we selected samples with quantified target outcomes and available chronological age and sex data. Models were generated using the *glmnet* package in R (version 4.1-8). Gaussian penalized regression models were trained on CpG site data and metabolite, protein, or clinical outcome levels using four alpha values (0.01, 0.1, 0.5, and 1) to assess the impact of varying the L1-L2 penalty balance. Hyperparameters were tuned through 10-fold cross-validation with *cv.glmnet* to select the optimal lambda for each alpha, minimizing mean squared error (MSE). All CpG sites and covariates were penalized, and features with at least five non-zero coefficients were retained in the final predictor models. Performance metrics—including MSE, mean absolute error (MAE), and Spearman correlation—were computed on test datasets to assess model fidelity. If more than one elastic net model was retained for a unique biomarker, we selected the one with the lowest MSE. Finally, we only selected those EBPs with Spearman test correlation higher than 0.2 ($p<0.05$).

XGBoost Training

XGBoost models were trained and evaluated in a high-throughput R pipeline (version 4.1.1) using the *data.table*, *xgboost*, and *caret* packages, with the goal of maximizing predictive accuracy through optimized hyperparameters. Pre-filtered mutual information (MI)-based datasets were generated. Hyperparameter tuning was conducted using a grid search on a parameter grid with *nrounds* set to 10, *eta* values of 0.05 and 0.2, alpha values of 0.1, 0.5, and 1, and lambda set to 1. The *caret* package's *train* function facilitated 3-fold cross-validation with parallel processing for efficiency. The optimal parameters identified from this grid search were applied to train the final XGBoost model, increasing *nrounds* to 1000 and incorporating an *early_stopping_rounds* parameter of 50 to prevent overfitting. This feature selection process improved model performance by reducing noise and enhancing signal quality, enabling the models to focus on CpG sites with the highest predictive potential for each target outcome. Model performance on the test data was assessed via mean squared error (MSE), mean absolute error (MAE), and Spearman correlation, with SHAP values calculated to capture CpG site contributions. We only selected those EBPs with Spearman test correlation higher than 0.2 ($p<0.05$) and without elastic net model retained.

Statistical analyses

Evaluation of EBP performance

To evaluate the performance of the EBPs, we calculated the Spearman correlation between the EBPs and the observed clinical, metabolite, and protein levels. We also compared the EBPs with previously developed DNAm surrogates, including 109 Protein EpiScores[11], 64 Nightingale metabolite surrogates[17], and 12 principal-component (PC) based surrogate proteins[3,18]. Of these, there was an overlap between EBPs and 18 Episcores, 6 Nightingale metabolites, and 5 common PC-based protein surrogates. Using the testing dataset, we calculated the correlations between all DNAm surrogates and the observed values.

Gene ontology enrichment analysis for clinical EBPs

To better understand the functional relevance of the CpG sites selected for the clinical models, we performed a gene ontology (GO) enrichment analysis, a widely used method to specify molecular function, cellular localization, and biological processes [41,42]. For each clinical outcome, we categorized CpG sites into positive and negative weights and performed two GO analyses using the top 5% of each category. We used the package *rGREAT* from R.

Association of EBPs to disease outcomes

To evaluate the clinical utility of EBPs, we analyzed their associations with diseases and compared these associations to those of the observed targeted measures. Diagnosis information, sourced from EMR data, allowed us to curate major chronic diseases prevalent in the MGB-ABC cohort using ICD-9/10-CM codes. For each disease, we identified both prevalent and incident cases by extracting the date of the first diagnosis. Patients diagnosed with a specific disease before plasma collection were classified as prevalent cases and included in a cross-sectional analysis. In this analysis, we estimated the odds ratio (OR) of each EBP or observed measure using logistic regression. Conversely, patients diagnosed after plasma collection were classified as incident cases and included in a prospective analysis, where we estimated the hazard ratio (HR) using a Cox proportional hazards model. All models were adjusted for age and sex to account for potential confounding factors. Additionally, we calculated the false discovery rate (FDR) to correct for inflated type I errors due to multiple comparisons.

Accuracy assessment of clinical EBP to detect deviations from recommended clinical ranges

The observed measurements, and EBP values of each biomarker were mean centered and normalized by dividing the individual values by the cohort's standard deviation. To transform the reference ranges accordingly, we subtracted from each one the mean of the corresponding observed value, and then divided the result by the observed standard deviation. The cases when both the EBP and the observed value are within the reference ranges (true positive), or outside of the reference ranges (true negative) are considered correct detections. EBPs outside of the reference range that correspond to within range observed values are called false positives, while EBPs within the reference range corresponding to observed values outside of range are considered false negatives. The accuracy is calculated as the percentage of correct detections with respect to the total number of measurements. The specificity is calculated as the number of true negatives divided by the total number of in range observed measurements, and the sensitivity is calculated as the number of true positives divided by the total number of out of range observed measurements.To identify the top performers, we used Youden's J-index, which is calculated as a sum of sensitivity and specificity subtracted from one. Biomarkers with higher J-index were considered top performers.

Longitudinal analysis

To evaluate the longitudinal changes, we considered 10 individuals with DNA methylation data and lab-based clinical biomarkers in at least timepoints from the TruDiagnostic cohort. The observed measurements, and EBP values of each biomarker were mean centered and normalized by dividing the result by the cohort’s standard deviation. We then calculated the difference between the EBP and the observed measurements for every time point, biomarker, and individual, and obtained the absolute value. Finally, for every biomarker and individual, we calculated the variance across the differences at the available timepoints.

Metadata analysis

The TruDiagnostic Cohort represents individuals who have performed TruDiagnostic testing commercially in the past and consented to participate in research. Due to the high cost of this testing and the reporting offering (Biological Aging), patients within this cohort are generally self-selected to be more affluent and healthier than the general population. When completing the test, users will fill out basic demographic, lifestyle, medication, and disease surveys. In order to analyze the association of lifestyle and demographic features to the EBPs, a team of clinicians selected EBPs and self-reported survey responses that were most likely to be linked. All the associations were evaluated using linear regression and adjusted by sex. Also, we evaluated these associations for each sex independently.

Author contributions

Data integrity: QC and VBD
Conceived and designed the study - JLS, RS, VBD
Performed analysis of physical samples for TruD cohort - TM
Data processing and normalization - QC, VBD, KS
Algorithm development - QC, VBD, DP, SV, SH
Statistical analysis and validation - QC, VBD, AA, NCG
Analyses and visualizations - LBD, RD, AC, TG
Drafted and edited manuscript - all authors

Data Availability

Requests for raw data, analyzed data and materials used to generate the results in this study will be reviewed by the contact PIs (Jessica Lasky-Su or Varun Dwaraka) to determine if the request is subject to intellectual property or confidentiality obligations. Datasets from electronic medical health records are not publicly available due to reasons of sensitivity. Restrictions apply to the use of these data due to privacy. Data that can be shared for research purposes upon request via a Data Use Agreement with the corresponding authors (Jessica Lasky-Su or Varun Dwaraka) by email (rejas@channing.harvard.edu or varun.dwaraka@trudiagnostic.com). Code to calculate all algorithms will be accessible via TruDiagnostic's DNAm Analysis Software after publication. You can request access to the software at https://www.trudiagnostic.com/softwarerequest.

Conflicts of interest

NC, LBD, IG, RD, DP, AC, SV, SH, TM, RS, and VBD are employees of TruDiagnostic. JLS is a scientific advisor to TruDiagnostic, Precision Inc and Ahara Inc. This work was completed under a sponsored research agreement between Brigham Women's Hospital and TruDiagnostic.

Ethical Statement

This study was conducted in accordance with all applicable ethical guidelines and regulations. Ethical approval was obtained from the Institutional Review Board (IRB)/Ethics Committee at BWH (2014P001109), and informed consent was obtained from all participants prior to their inclusion in the study. Data were anonymized to protect participant privacy, and confidentiality was strictly maintained throughout the study. All human subject research adhered to the principles outlined in the Declaration of Helsinki, ensuring respect for individuals, beneficence, and justice.

Funding

Efforts for QC, YC, TG, and JALS are supported by R01HL123915 and R01HL155742. Efforts for AA, JALS, and KM are supported by OT2HL161841 from the NIH/NHLBI. Efforts for YC and JALS are supported by R01HL141826 from the NIH/NHLBI. Efforts for JALS are supported by R01HL169300 from the NIH/NHLBI. Efforts for JALS are supported by U19AI168643 from the NIH/NIAID. K.S. is supported by the Biomedical Research Program at Weill Cornell Medicine in Qatar, a program funded by the Qatar Foundation, and also supported by Qatar National Research Fund (QNRF) grants NPRP11C-0115-180010 and ARG01-0420-230007.

Conflicts of interest. This work was funded in part by TruDiagnostic. JLS is a scientific advisor to Precion Inc and TruDiagnostic Inc. JLS, QC, VBD, RS have filed patents on work from this publication.

REFERENCES

1. Alegría-Torres, J. A., Baccarelli, A. & Bollati, V. Epigenetics and lifestyle. *Epigenomics* **3**, 267 (2011).
2. Heijmans, B. T. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci.* **105**, 17046–17049 (2008).
3. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, 303–327 (2019).
4. Hillary, R. F. *et al.* Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden. *Clin. Epigenetics* **12**, 115 (2020).
5. Zhang, Y. *et al.* DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat. Commun.* **8**, 14617 (2017).
6. Jones, A. C. *et al.* A methylation risk score for chronic kidney disease: a HyperGEN study. *Sci. Rep.* **14**, 17757 (2024).
7. Thompson, M. *et al.* Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. *Npj Genomic Med.* **7**, 1–11 (2022).
8. Stevenson, A. J. *et al.* Creating and Validating a DNA Methylation-Based Proxy for Interleukin-6. *J. Gerontol. Ser. A* **76**, 2284–2292 (2021).
9. Zaghlool, S. B. *et al.* Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. *Nat. Commun.* **11**, 15 (2020).
10. Conole, E. L. *et al.* DNA Methylation and Protein Markers of Chronic Inflammation and Their Associations With Brain and Cognitive Aging. *Neurology* **97**, e2340 (2021).
11. Gadd, D. A. *et al.* Epigenetic scores for the circulating proteome as tools for disease prediction. *eLife* **11**, e71802 (2022).
12. van den Akker, E. B. *et al.* Metabolic Age Based on the BBMRI-NL 1H-NMR Metabolomics Repository as Biomarker of Age-related Disease. *Circ. Genomic Precis. Med.* **13**, 541–547 (2020).

13. Chen, Q. *et al.* OMICmAge: An integrative multi-omics approach to quantify biological age with electronic medical records. 2023.10.16.562114 Preprint at https://doi.org/10.1101/2023.10.16.562114 (2023).

14. Lu, A. T. *et al.* DNA methylation GrimAge version 2. *Aging* **14**, 9484–9549 (2022).

15. Sadikovic, B. *et al.* Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. *Genet. Med.* **23**, 1065–1074 (2021).

16. Nalichowski, R., Keogh, D., Chueh, H. C. & Murphy, S. N. Calculating the Benefits of a Research Patient Data Repository. *AMIA. Annu. Symp. Proc.* **2006**, 1044 (2006).

17. Bizzarri, D. *et al.* NMR metabolomics-guided DNA methylation mortality predictors. *eBioMedicine* **107**, 105279 (2024).

18. Higgins-Chen, A. T. *et al.* A computational solution for bolstering reliability of epigenetic clocks: implications for clinical trials and longitudinal tracking. *Nat. Aging* **2**, 644–661 (2022).

19. Green, E., Todd, B. & Health, D. Mechanism of glucocorticoid regulation of alkaline phosphatase gene expression in osteoblast-like cells. *Eur. J. Biochem.* **188**, 147–153 (1990).

20. Sowa, H., Kaji, H., Yamaguchi, T., Sugimoto, T. & Chihara, K. Smad3 Promotes Alkaline Phosphatase Activity and Mineralization of Osteoblastic MC3T3-E1 Cells. *J. Bone Miner. Res.* **17**, 1190–1199 (2002).

21. Tokunaga, A. *et al.* Selective inhibition of low-affinity memory CD8+ T cells by corticosteroids. *J. Exp. Med.* **216**, 2701–2713 (2019).

22. Konishi, A. *et al.* Glucocorticoid imprints a low glucose metabolism onto CD8 T cells and induces the persistent suppression of the immune response. *Biochem. Biophys. Res. Commun.* **588**, 34–40 (2022).

23. Schwartze, J. T. *et al.* Glucocorticoids Recruit Tgfbr3 and Smad1 to Shift Transforming Growth Factor-β Signaling from the Tgfbr1/Smad2/3 Axis to the Acvrl1/Smad1 Axis in Lung Fibroblasts. *J. Biol. Chem.* **289**, 3262 (2013).

24. Gadd, D. A. *et al.* DNAm scores for serum GDF15 and NT-proBNP levels associate with

a range of traits affecting the body and brain. 2023.10.18.23297200 Preprint at https://doi.org/10.1101/2023.10.18.23297200 (2023).

25. Stevenson, A. J. *et al.* Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clin. Epigenetics* **12**, 113 (2020).

26. Hillary, R. F. *et al.* Blood-based epigenome–wide analyses of chronic low–grade inflammation across diverse population cohorts. 2023.11.02.23298000 Preprint at https://doi.org/10.1101/2023.11.02.23298000 (2023).

27. CDC. Chronic Kidney Disease in the United States, 2023. *Chronic Kidney Disease* https://www.cdc.gov/kidney-disease/php/data-research/index.html (2024).

28. Schrader, S. *et al.* Novel Subgroups of Type 2 Diabetes Display Different Epigenetic Patterns That Associate With Future Diabetic Complications. *Diabetes Care* **45**, 1621–1630 (2022).

29. Ho, A. *et al.* Circulating glucuronic acid predicts healthspan and longevity in humans and mice. *Aging* **11**, 7694–7706 (2019).

30. Keszthelyi, T. M. & Tory, K. The importance of pseudouridylation: human disorders related to the fifth nucleoside. *Biol. Futura* **74**, 3–15 (2023).

31. Guillen-Angel, M. & Roignant, J.-Y. Exploring pseudouridylation: dysregulation in disease and therapeutic potential. *Curr. Opin. Genet. Dev.* **87**, 102210 (2024).

32. di Candia, A. M., de Avila, D. X., Moreira, G. R., Villacorta, H. & Maisel, A. S. Growth differentiation factor-15, a novel systemic biomarker of oxidative stress, inflammation, and cellular aging: Potential role in cardiovascular diseases. *Am. Heart J. Plus Cardiol. Res. Pract.* **9**, 100046 (2021).

33. Kang, J. *et al.* Ribosomal proteins and human diseases: molecular mechanisms and targeted therapy. *Signal Transduct. Target. Ther.* **6**, 1–22 (2021).

34. Li, J. *et al.* Overview of growth differentiation factor 15 (GDF15) in metabolic diseases. *Biomed. Pharmacother.* **176**, 116809 (2024).

35. Castro, V. M. *et al.* The Mass General Brigham Biobank Portal: an i2b2-based data repository linking disparate and high-dimensional patient data to support multimodal analytics. *J. Am. Med. Inform. Assoc. JAMIA* **29**, 643 (2021).

36. Murphy, S. N. *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc. JAMIA* **17**, 124 (2010).

37. Foox, J. *et al.* The SEQC2 epigenomics quality control (EpiQC) study. *Genome Biol.* **22**, 332 (2021).

38. Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M. & Milgram, E. Integrated, Nontargeted Ultrahigh Performance Liquid Chromatography/Electrospray Ionization Tandem Mass Spectrometry Platform for the Identification and Relative Quantification of the Small-Molecule Complement of Biological Systems. *Anal. Chem.* **81**, 6656–6667 (2009).

39. Sha, W. *et al.* Metabolomic profiling can predict which humans will develop liver dysfunction when deprived of dietary choline. *FASEB J.* **24**, 2962–2975 (2010).

40. Suhre, K. *et al.* A genome-wide association study of mass spectrometry proteomics using the Seer Proteograph platform. 2024.05.27.596028 Preprint at https://doi.org/10.1101/2024.05.27.596028 (2024).

41. Consortium, T. G. O. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000).

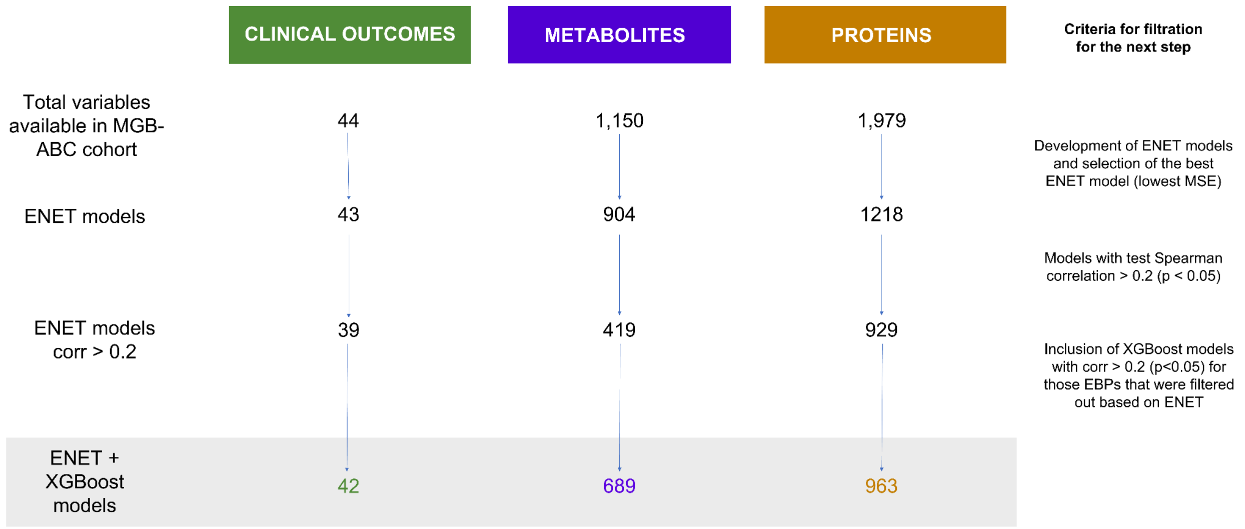
42. Tomczak, A. *et al.* Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci. Rep.* **8**, 5115 (2018).

Figures and Tables

A

	MGB - Aging Biobank Cohort	**TruDiagnostic Biobank**
Samples	4,418	31,012
Unique individuals	4,418	26,155
Sex, male, N (%)	1,566 (35.4%)	15288 (58.5%)
Age in years, mean (range)	56.5 (18.3, 96.3)	52.0 (4.4-100.0)
Race/Ethnicity, N (%)		
White	3,503 (79.3%)	19957 (76.3%)
African American	354 (8.0%)	1094 (4.2%)
Asian	85 (1.9%)	2353 (9.0%)
Other	353 (8.0%)	1788 (6.8%)
Unknown/Missing	123 (2.8%)	963 (3.7%)

B



	CLINICAL OUTCOMES	METABOLITES	PROTEINS	Criteria for filtration for the next step
Total variables available in MGB-ABC cohort	44	1,150	1,979	Development of ENET models and selection of the best ENET model (lowest MSE)
ENET models	43	904	1218	Models with test Spearman correlation > 0.2 (p < 0.05)
ENET models corr > 0.2	39	419	929	Inclusion of XGBoost models with corr > 0.2 (p<0.05) for those EBPs that were filtered out based on ENET
ENET + XGBoost models	42	689	963	

Figure 1. Demographics and flowchart showing the development and selection of the Epigenetic Biomarker Proxies (EBPs). (**A**) Demographics of the cohorts used in the EBP development and validation. (**B**) Flowchart showing the selection of the EBPs. At the top, the total number of biomarkers available in the Massachusetts General Brigham's Aging Biobank Cohort (MGB-ABC). At the bottom, the total number of EBPs selected.

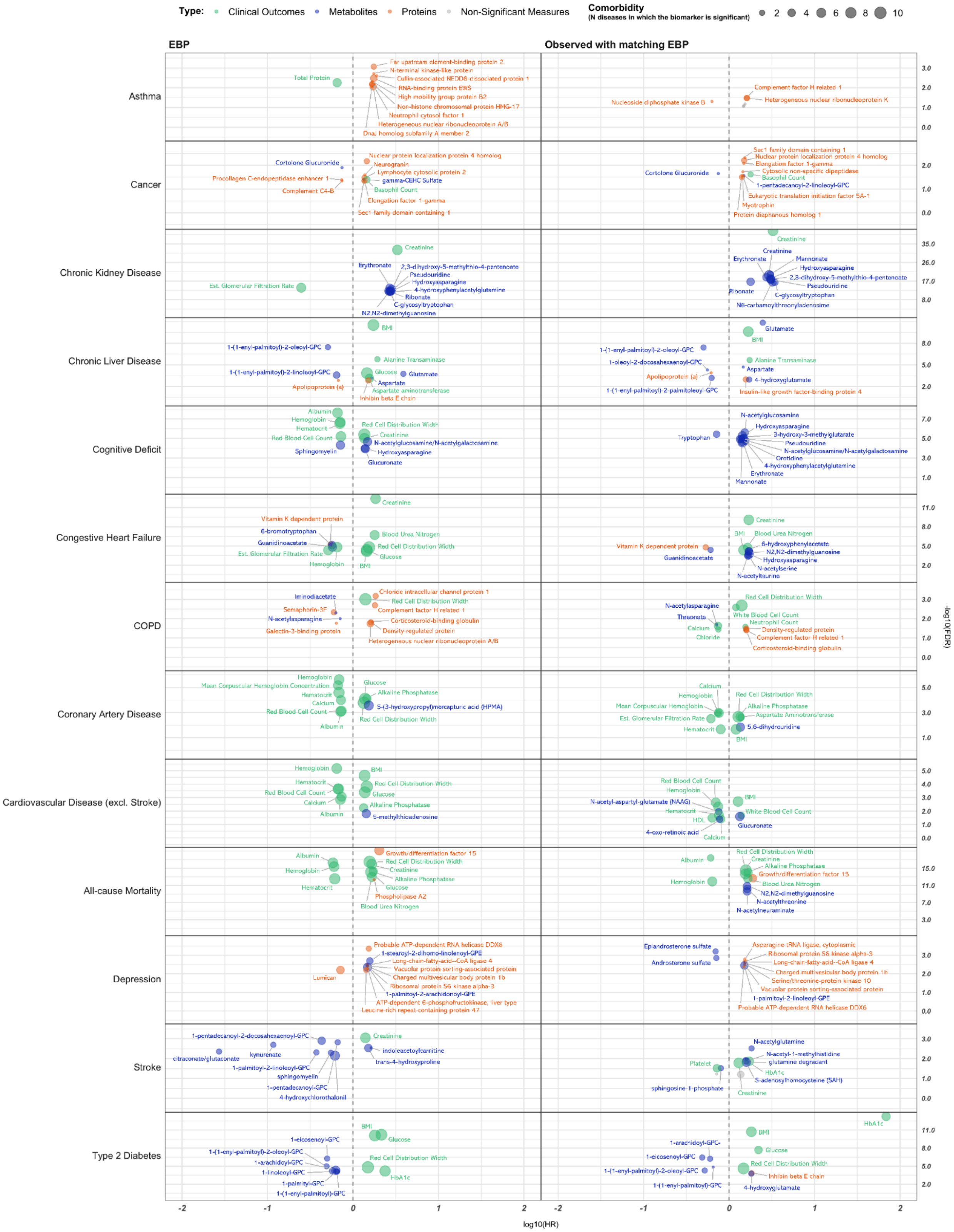


Figure 2. Summary volcano plot. Volcano plot showing the top 10 EBPs and observed value for 13 diseases based on Odds Ratio.

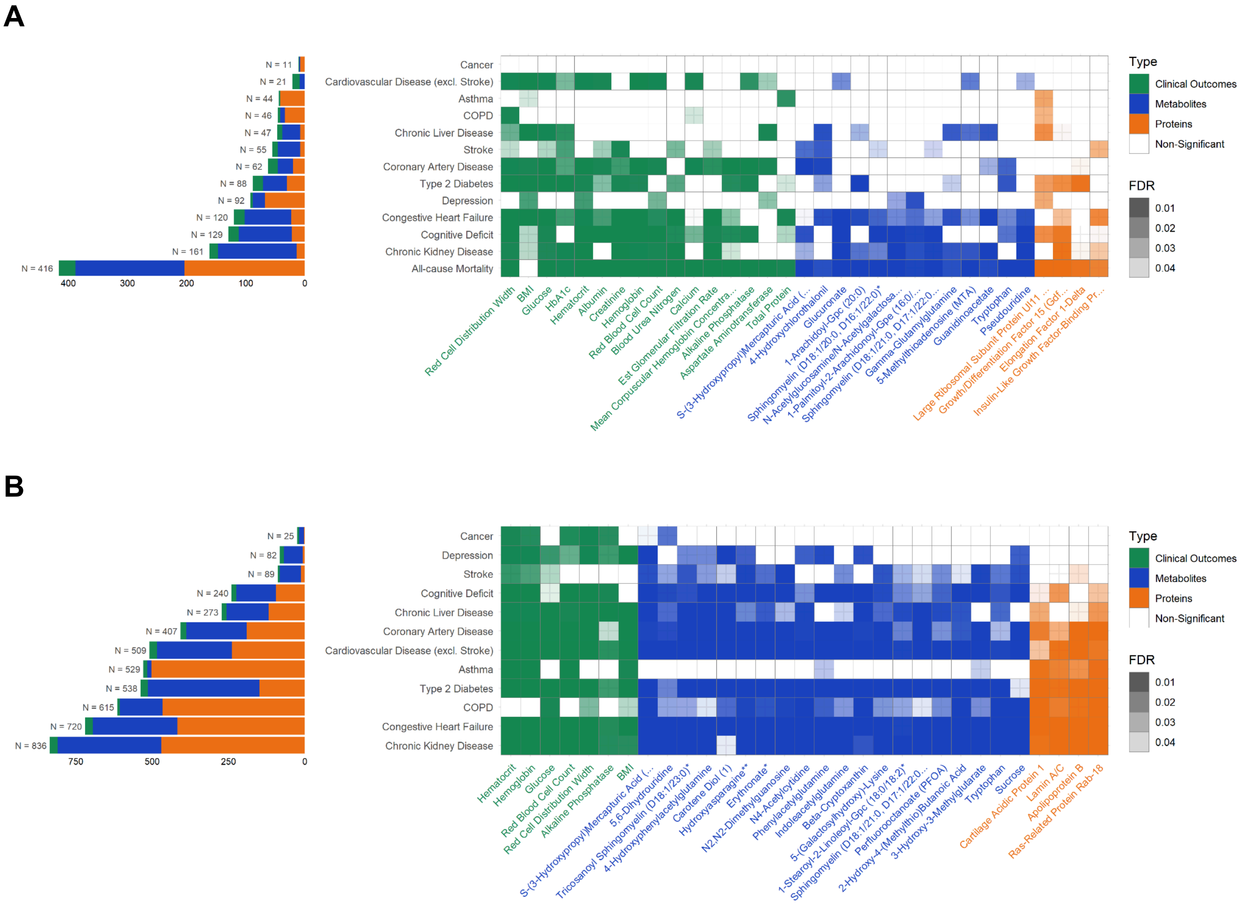


Figure 3. Heatmap showing the significant EBPs for each disease and the EBPs associated with 5 or more diseases. (**A**) The plot is according to incident associations, which are based on Hazard Ratios (HRs). The EBPs associated with 5 or more diseases are included. (**B**) The plot is according to prevalent associations, which are based on Odds Ratios (ORs). The EBPs associated with 9 or more diseases are included. The horizontal bars represent the number of EBPs with FDR < 0.05 for each disease. In green, the clinical EBPs; in blue, the metabolite EBPs; and in orange, the protein EBPs. In the heatmap, the darkness of the squares is proportional to the FDR, being the lightest 0.05 (non-significant) and the darkest 1.

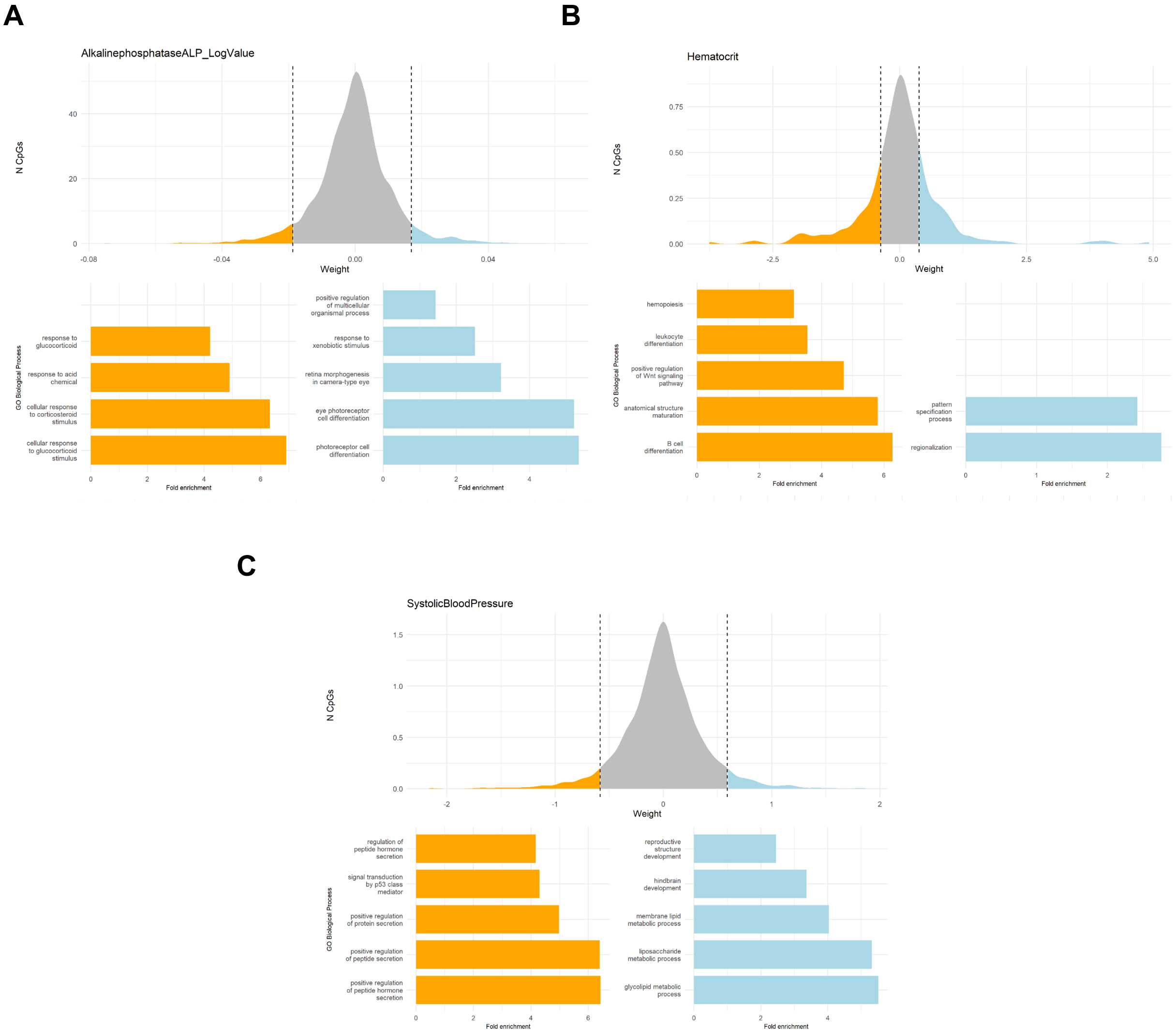


Figure 4. Gene ontology enrichment of strongest weighted features of EBPs. (A) GO enrichment for alkaline phosphatase, **(B)** GO enrichment for hematocrit, **(C)** GO enrichment for systolic blood pressure. For each clinical outcome, we categorized CpG sites into positive weights and negative weights and performed two GO analyses using the top 5% of each category. In the case of the hematocrit, we subset the 20% to have at least 50 CpGs per category.

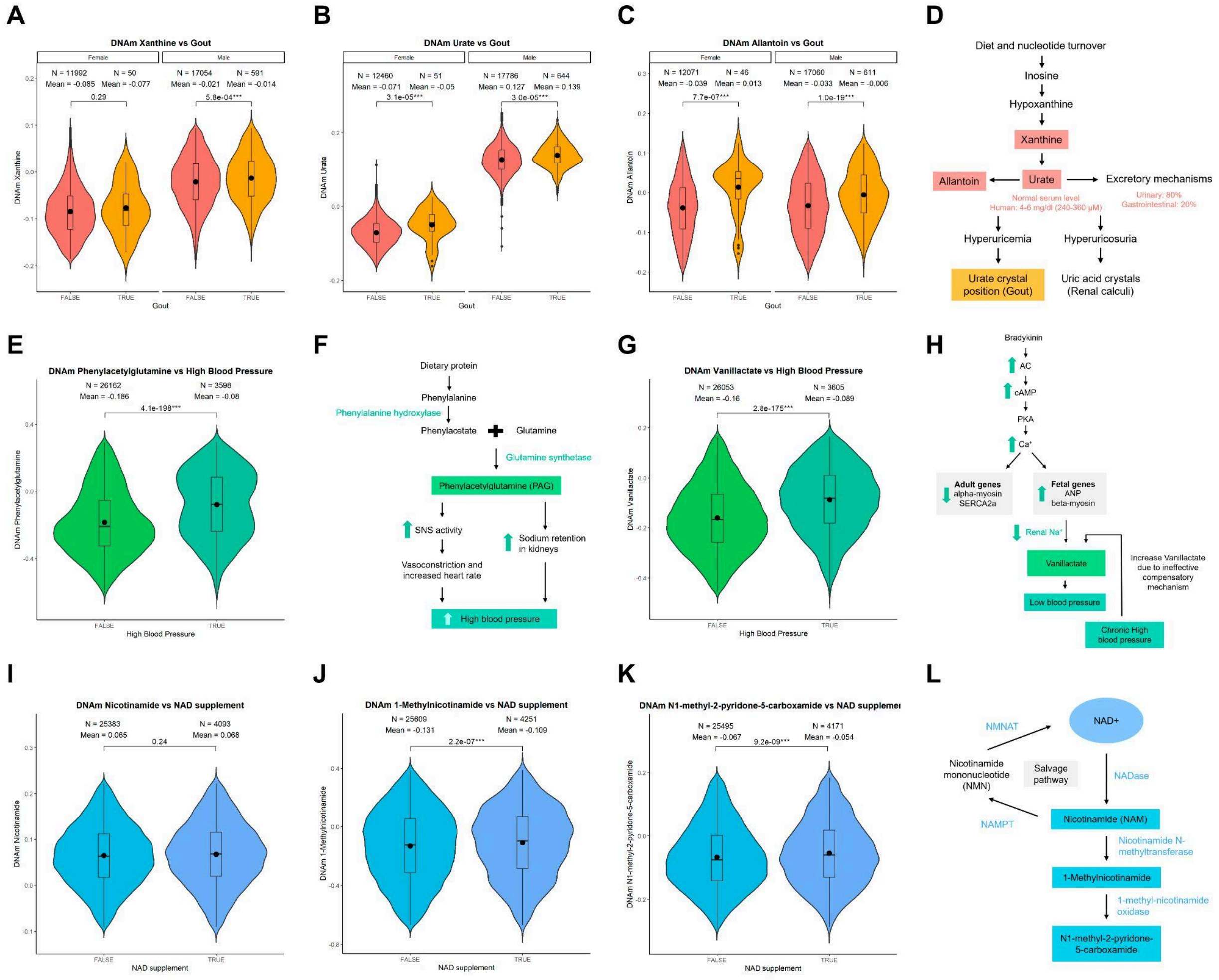


Figure 5. Validation analyses evaluating the changes of the EBP levels in different scenarios. (**A-C**) Association of Xanthine, Urate, and Allantion EBPs with the presence or absence of gout. (**D**) Uric acid pathway. (**E**) Association of Phenylacetylglutamine EBP with high blood pressure. (**F**) Implication of Phenylacetylglutamine in the high blood pressure prevalence. (**G**) Association of Vanillactate EBP with high blood pressure. (**H**) Implication of Vanillactate in the high blood pressure prevalence. (**I-K**) Association of Nicotinamide, 1-Methylnicotinamide, and N1-methyl-2-pyridone-5-carboxamide EBPs with NAD Supplementation. (**L**) NAD supplementation mode of action.

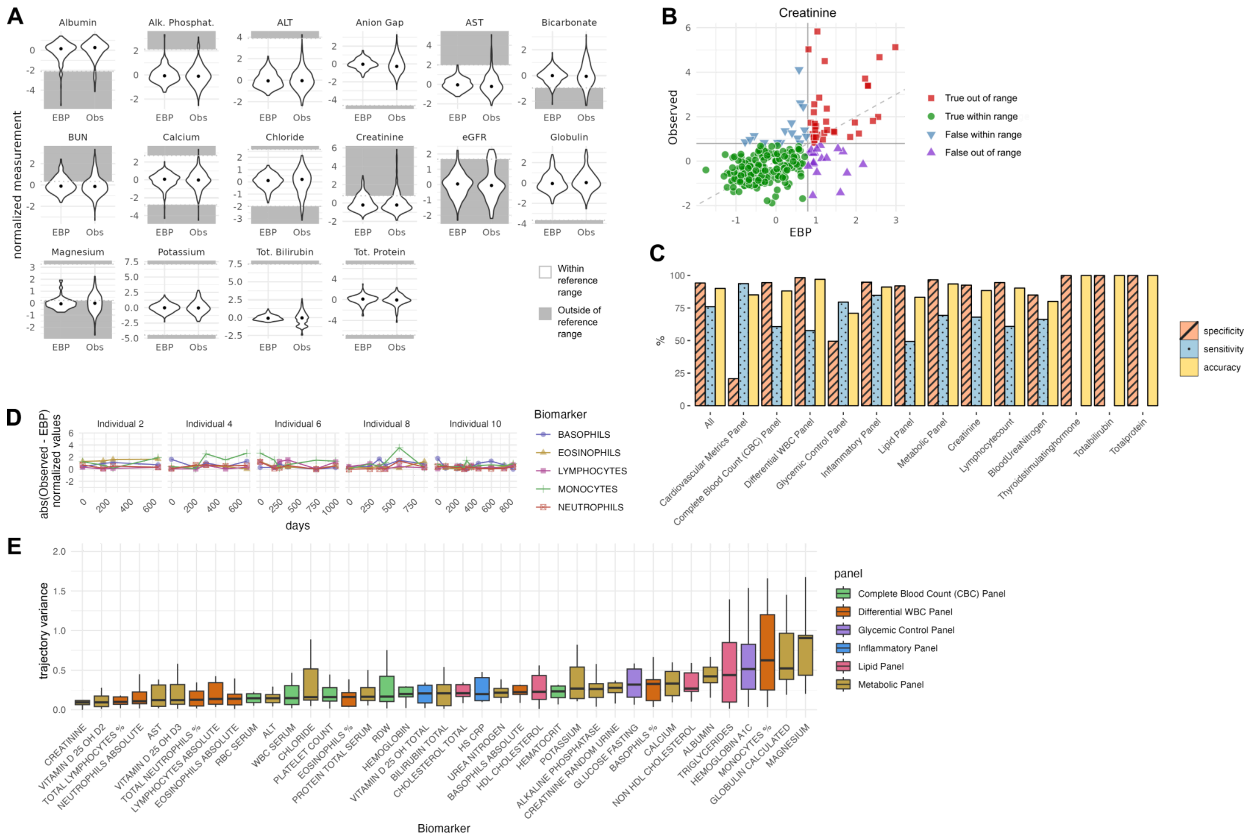


Figure 6. Analysis comparing the observed values to the EBP to assess accuracy of longitudinal tracking. The single clinical warnings detection analysis was done using the test set of the MGB cohort. The longitudinal analysis was done using data from 9 individuals in the TruD cohort having 4 or more samples with matched DNA methylation and clinical data. (**A**) EBPs and observed biomarkers with respect to clinical reference ranges. (**B**) EBP vs Observed measurements of Creatinine with respect to the reference range. (**C**) Accuracy, sensitivity, and specificity metrics of EBPs at detecting within range or out of range observed values. (**D**) Absolute difference between EBPs and observed values at available time points for 5 individuals in the TruD cohort. (**E**) Variance in EBP-Observed values difference across time points for all clinical biomarkers.

	HR				OR			
	Total Significant OBS	**Total Significant EBP**	**Shared**	**Shared in which EBP outperforms OBS**	**Total Significant OBS**	**Total Significant EBP**	**Shared**	**Shared in which EBP outperforms OBS**
All-cause Mortality	**346**	**416**	**301**	238(79%)	NA	NA	NA	NA
Asthma	**3**	**44**	**3**	2(67%)	**368**	**529**	**357**	343(96%)
Cancer	**26**	**11**	**9**	1(11%)	**11**	**25**	**8**	4(11%)
Chronic Kidney Disease	**206**	**161**	**126**	58(46%)	**796**	**836**	**676**	427(46%)
Chronic Liver Disease	**51**	**47**	**32**	18(56%)	**362**	**273**	**227**	102(56%)
COPD	**11**	**46**	**9**	4(44%)	**457**	**615**	**400**	341(44%)
CVD (excl. Stroke)	**11**	**21**	**8**	7(88%)	**484**	**509**	**380**	206(88%)
Congestive Heart Failure	**94**	**120**	**70**	44(63%)	**689**	**720**	**572**	354(63%)
Coronary Artery Disease	**16**	**62**	**14**	13(93%)	**407**	**407**	**297**	136(93%)
Depression	**82**	**92**	**57**	12(21%)	**117**	**82**	**61**	28(21%)
Stroke	**8**	**55**	**4**	3(75%)	**94**	**89**	**57**	29(75%)
Type 2 Diabetes	**87**	**88**	**52**	28(54%)	**483**	**538**	**397**	249(54%)
Cognitive Deficit	**122**	**129**	**84**	49(58%)	**227**	**240**	**162**	90(58%)
Total	**1063**	**1292**	**769**	**477(62%)**	**4495**	**4863**	**3594**	**2309(64%)**
Unique	**597**	**703**	**457**	**332(73%)**	**1436**	**1475**	**1256**	**1056(85%)**

Significant: FDR<0.05.
EBP: Epigenetic Biomarker Proxy.
OBS: Observed value.

Table 1. Summary of the significant Hazard Ratios (HR) and Odd Ratios (OR) for different chronic diseases. HRs and ORs values for 13 chronic diseases comparing the Epigenetic Biomarker Proxies (EBPs) against the actual observed values.