An Objective Metric of Individual Health And Aging For Population Surveys

Qing Li
Xinjiang University
Véronique Legault
University of Sherbrooke
Vincent-Daniel Girard
University of Sherbrooke
Luigi Ferrucci
NIA: National Institute on Aging
Linda P. Fried
Columbia University
Alan A Cohen (alan.cohen@usherbrooke.ca)
University of Sherbrooke https://orcid.org/0000-0003-4113-3988

Research

Keywords: Physiological dysregulation, biomarkers, Mahalanobis distance, population, composition, allostatic load, self-assessed health

Posted Date: October 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-952500/v1

Version of Record: A version of this preprint was published at Population Health Metrics on March 31st, 2022. See the published version at https://doi.org/10.1186/s12963-022-00289-0.

An objective metric of individual health and aging for population surveys

Qing Li[1], Véronique Legault[2], Vincent-Daniel Girard[2], Luigi Ferrucci[3], Linda P. Fried[4], Alan A. Cohen[2,5,6]*,

[1] School of Economics and Management, Xinjiang University, 666 Shengli Road, Urumqi, 830046, China.

[2] PRIMUS Research Group, Department of Family Medicine, University of Sherbrooke, 3001 12e Ave N, Sherbrooke, QC, J1H 5N4, Canada.

[3] Translational Gerontology Branch, Longitudinal Studies Section, National Institute on Aging, National Institutes of Health, MedStar Harbor Hospital, 3001 S. Hanover Street, Baltimore, MD, 21225, United States of America.

[4] Mailman School of Public Health, Columbia University, 722 W. 168th Street, R1408, New York, NY, 10032, United States of America.

[5] Research Center on Aging, 1036 Belvédère S, Sherbrooke, QC, J1H 4C4, Canada.

[6] Research Center of Centre Hospitalier Universitaire de Sherbrooke, 3001 12e Ave N, Sherbrooke, QC, J1H 5N4, Canada.

* Correspondence to: Alan.Cohen@USherbrooke.ca.

Abstract

Background: Generalized, biomarker-based metrics of health status have numerous applications in fields ranging from sociology and economics to clinical research. We recently proposed a novel metric of health status based on physiological dysregulation measured as a Mahalanobis distance (DM) among clinical biomarkers. While DM was not particularly sensitive to the choice of biomarkers, it required calibration when used in different populations, making it difficult to compare findings across studies. To facilitate its use, here we aimed to identify and validate a standard version of DM that would be highly stable across populations, while using fewer biomarkers drawn exclusively from common blood panels. **Methods**: Using three datasets, we identified nine-biomarker (DM9) and seventeen-biomarker (DM17) versions of DM, choosing biomarkers based on their consistent levels across populations. We validated them in a fourth dataset. We assessed DM stability within and across populations by looking at correlations of DM versions calibrated using different populations or their demographic subsets. We used regression models to compare these standard DM versions to allostatic load and self-assessed health in their association with diverse health outcomes. **Results**: DM9 and DM17 were highly stable across population subsets (mean $r = 0.96$ and 0.95, respectively) and across populations (mean $r = 0.94$ for both). Performance predicting health outcomes was competitive with allostatic load and self-assessed health, though performance of these markers were somewhat variable for different health outcomes. **Conclusions**: Both DM9 and DM17 are highly stable within and across populations, supporting their use as objective metrics of health status. DM17 performs slightly better than DM9 and at least as well as other comparable metrics, but requires more biomarkers. The metrics we propose here are easy to measure with data that are available in a wide array of panel, cohort, and clinical studies.

Keywords: Physiological dysregulation, biomarkers, Mahalanobis distance, population composition, allostatic load, self-assessed health

Background

A key challenge in the study of population health is the operationalization of a metric for global health status. In addition to potential clinical use at the individual level, such a metric would serve many purposes at the population level. It could serve as a control/adjustment variable, similar to how socioeconomic status and age are adjusted for in many epidemiological studies. It could serve as a short-term or intermediate outcome for interventions, either clinical or policy. It could be used by diverse fields ranging from health economics to sociology, demography, epidemiology, and clinical research. One approach to this problem has been using subjective metrics of global health such as self-reported health. However, subjective perception of health is conditioned by cultural or social norms as well as by medical diagnosis and access to health-care resources [1]. Thus, unless a subjective component is a main dimension to be addressed, it may be preferable to use objective health metrics that tend to be more stable [2], although specific criteria for their construction is still a matter of discussion.

A major challenge is that health is unquestionably multidimensional, and summarizing information from different indicators into a single index is not a straightforward problem. Defining the dimensions is challenging and has not yet been the study of rigorous study, to our knowledge. Various metrics of comorbidity, multimorbidity and frailty have been

proposed in the literature [3–5], though most of them show limited variation among healthy younger and middle-aged adults because they are based on elements that only occur late in life. In this context, the deficit accumulation approach to frailty, based on a simple count of potential health deficits present in an individual, is particularly attractive because it is relatively robust to the precise choice of deficits in the list and health deficits can thus identify a wide range of severities, some of which are manifested even in younger individuals [6,7]. But despite the wide use of metrics based on deficit accumulation, a standardized version has yet to be developed [8].

On the other hand, there are biomarker-based metrics that attempt to integrate the signal of multiple aspects of health. Perhaps the best-known of these is allostatic load [9]. Allostatic load is based on the theory that chronic stress can leave physiological sequalae that can be measured by creating a metric of common biomarkers linked to appropriate physiological systems: neuro-endocrine stress (cortisol, epinephrine, norepinephrine), metabolic markers (blood pressure, lipid profiles, glucose metabolism, obesity metrics), as well as a few additional biomarkers (inflammatory markers, DHEA-S, IGF-1, etc.) [10,11]. However, allostatic load is challenging because it is conceptualized based on circular reasoning: the proxy metrics are chosen because of their known association with health and aging, so it is unsurprising the sum does as well [12]. Because it is often operationalized as a count of how many of the factors exceed clinical bounds, measures of allostatic load end up resembling comorbidity metrics in many ways, though the latter are generally not biomarker based.

Recently, our lab group has developed an alternative biomarker-based metric of physiological dysregulation based on a statistical distance (specifically, Mahalanobis distance) among biomarkers [13]. The idea is that a population average is an approximation of a homeostatic state, and that deviations from this multivariate biomarker average represent dysregulation and thus should increase with age and predict poor health state. Indeed, we have shown that dysregulation rates increase with age within individuals, and predict multiple health outcomes (mortality, frailty, various chronic diseases) after controlling for age [14,15]. A lack of sensitivity to precise biomarker choice, and an increasing signal with more biomarkers confirm a complex systems interpretation of dysregulation as an emergent property of physiological regulatory networks [16]. Results can be replicated in many human populations [13,14,17–27] and even in captive primates [28] and wild birds [29]. Lastly, dysregulation can be measured either globally or by specific physiological system [30], opening up the possibility for much more detailed characterization of health state.

The dysregulation approach presents a number of clear advantages. All variables are left continuous, so there is no information loss due to categorization. The scale from 0 to infinity is appropriate for measuring dysregulation. Because it uses distances from the mean of each biomarker rather than absolute levels, it agrees with theory on biological homeostasis, which suggests that intermediate values of individual biomarkers should generally be optimal, and with evidence that variance increases with dysregulation [31,32]. The Mahalanobis distance also incorporates the correlation structure of the variables, appropriately down-weighting redundancy among biomarkers. The insensitivity to biomarker choice means that it can be

117 easily applied in existing datasets, can be applied in clinical contexts, and can be applied
118 cheaply without requiring fancy, cutting-edge biomarkers. Importantly, it avoids the
119 circularity problems present with allostatic load and metrics of biological age: the
120 biomarkers are not selected based on correlations with age or health state, and there is no
121 required calibration with age or health state, so the signal is an independent indicator of
122 physiological state.
123
124 Nonetheless, some of these same advantages also present challenges. First, the possibility to
125 use nearly any broad combination of biomarkers means that there is no standard version, and
126 that values from one study cannot be compared directly to those from another. Second, while
127 the approach works in every human population tested, differences in biomarker levels and
128 correlations across populations mean that separate calibration (calculation of the mean vector
129 and variance-covariance matrix) is required for each population. This poses problems for
130 small studies (e.g. in clinical research) where the sample is too small to provide a robust
131 estimation of these parameters. Third, the combination of these issues means that there are
132 technical challenges for potential users who are less statistically inclined and would like a
133 simple recipe or automatic calculator.
134
135 Here, we present a standardized version of a biomarker-based global health metric that
136 overcomes these problems. Specifically, we provide a clear methodology and rationale for
137 choosing a subset of biomarkers that provide a strong signal, are readily available in most
138 contexts, and can be calibrated across populations, not just within. We demonstrate the
139 stability of the metric and its predictive power for health outcomes compared to self-reported
140 health and allostatic load. We call the metric "DSign" for Dysregulation Signature, and

propose a principle version based on 17 biomarkers and a secondary version based on 9 biomarkers, for cases in which all 17 may not be available. All biomarkers in both versions are standard clinical markers that can be readily measured in almost any setting for a very reasonable cost (e.g. <$1/marker).

Methods

Datasets

To construct our standard DM versions, we used data from two longitudinal cohort studies and one cross-sectional survey (see Table 1 for details): the Baltimore Longitudinal Study of Aging (BLSA), Invecchiare in Chianti (InCHIANTI), and the National Health and Nutrition Examination Survey (NHANES). BLSA, one of the world's longest studies of aging in humans, is composed of community-dwelling adults in the Baltimore and Washington DC areas aged 21–96 [33]. A 2003 re-design of methodology was tailored to improve the inference for systems-level questions [34], and we use data on 1205 individuals from after this date. InCHIANTI is a prospective population-based study of 1156 adults aged 65–102 and 299 aged 20–64, randomly selected using multistage stratified sampling from two towns in Tuscany, Italy [35]. NHANES is a continuous cross-sectional stratified survey designed to be representative of the US population. Data are updated approximately every year and are made available freely (Centers for Disease Control and Prevention of the U.S. Department of Health and Human Services; http://www.cdc.gov/nchs/nhanes.htm). We used individuals aged 20 years or older from the waves 1999–2000, 2001–2002, 2003–2004, 2005–2006, 2007–2008, and 2009–2010, which have been described in detail elsewhere [36].

163 Validation of our standard DM versions was performed with the Women's Health and
164 Aging Study (WHAS). WHAS is a population-based prospective study of community-
165 dwelling women drawn from eastern Baltimore City and Baltimore County, originally
166 consisting of two separate studies: WHAS I, which includes 1002 women aged 65+ among
167 the 1/3 most disabled in the population [37], and WHAS II, which includes 436 women aged
168 70–79 among the 2/3 least disabled [38].
169
170 ***Distance-based metric of physiological dysregulation***
171 To calculate our dysregulation score (DM), we consider individuals as points in a multi-
172 dimensional biomarker space, where each biomarker is an axis of the space. DM defines a
173 reference population (RP) whose centroid approximates "the ideal state", and then calculates
174 the Mahalanobis distance to the centroid for each individual, according to equation 1 [39]:

175 $$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \qquad (1)$$

176 where x is a vector of simultaneously observed values for the biomarkers, μ is the
177 equivalent-length vector of means for each biomarker in the RP, and ∑ is the variance-
178 covariance matrix of the biomarkers in the RP. Before DM calculation, all variables are
179 transformed as necessary (log or square root) to approach normality. For each biomarker, a
180 single best transformation was identified across datasets. The Mahalanobis distance can
181 become unreliable when the scales of the variables differ; we thus standardize each
182 biomarker with respect to the mean and standard deviation of the RP. Because it is
183 approximately log-normally distributed, we used the logarithm of DM in subsequent
184 analyses.

In calculating DM, we do not give any special weight to any of the biomarkers over and above the weights implicit in the covariance matrix. Although certain biomarkers are well-known to be important for certain diseases or physiological systems (e.g. glucose for diabetes), there is no consensus that one biomarker is more important for general health than another; subjective weighting of individual variables could thus introduce a bias in the metric.

Selection of biomarkers

Thirty-one biomarkers were available in all datasets in sufficient sample sizes (see Supplementary Table 1 and Supplementary Fig. 1 in Additional File 1): hemoglobin, hematocrit, red cell distribution width (RDW), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), red blood cell count (RBC), platelets, white blood cells (WBC), basophil percentage (BASO%), lymphocyte percentage (LYM%), monocyte percentage (MONO%), neutrophil percentage (NEUT%), ferritin, glucose, calcium, chloride, sodium, potassium, vitamin B12, folate, total cholesterol, triglycerides, high density lipoprotein (HDL), albumin, alkaline phosphatase (ALKP), total proteins, gamma-glutamyl transferase (GGT), lactate dehydrogenase (LDH), uric acid, alanine transaminase (ALT), aspartate transaminase (AST).

We used a multistep approach in order to select biomarkers that respond best to the following criteria: 1) stability of DM across various RPs; 2) biological signal, as measured by concordance with DM calculated using the full set (i.e. 31 biomarkers); 3) availability of biomarkers in clinical/research contexts; 4) diversity of physiological systems represented; 5) redundancy among biomarkers; and 6) consistency of mean biomarker levels within and across populations. The detailed approach that led to our final sets of biomarkers can be

found in Additional File 1. Those final sets are the following: a 9-biomarker set (DM9) composed of MCH, RDW, platelets, RBC, hemoglobin, WBC, BASO%, HDL, and LYM%; and a 17-biomarker set (DM17) composed of the same biomarkers as DM9, but also including GGT, AST, ALKP, albumin, total proteins, calcium, potassium, and vitamin B12.

Stability of final DM sets according to the choice of the reference population

An important issue regarding DM calculation is the choice of the RP, which is used to calculate the centroid (the biomarker combination assumed to represent optimal health) as well as the variance-covariance matrix. While there are widely accepted normal ranges for individual biomarkers, there is no consensus on a point-wise multivariate centroid that represents optimal health status. Our previous work suggested that, while a younger and healthier RP could yield a slightly better signal, the RP should not be too demographically different from the study population [17]. The entire study population itself is generally a good approximation. While it might seem intuitive that the mean of a younger, healthier population should provide a better estimate of optimal state, many age-related changes in biomarkers may actually be compensations to other changes [40,41], raising the possibility that age-specific RPs could be preferable. Since we cannot separate out pathological changes from compensatory changes, in practice all RPs will confound these two effects to some extent.

Our goal here was to choose biomarkers that are less sensitive to the choice of RP in order to eliminate the need to consider all these factors, and to facilitate the use of a single RP for nearly any study in nearly any context. We assessed the stability of our final sets within and across populations. To test for stability across RPs, we computed correlations between DM calculated using the study population as its own RP, or using another dataset as

the RP. Large numbers of such analyses were compiled in correlation matrices, and correlation coefficients were averaged. To test for stability within populations, we divided each dataset into subsets (by sex, age, race, education level, or marital status) and performed similar correlations, i.e. between DM calculated using a given subset as its own RP or using another subset.

Association of final biomarker sets with mortality, frailty, and comorbidities

To assess whether our final sets of biomarkers are truly representative of physiological dysregulation, we explored their association with mortality, clinical frailty, cardiovascular disease (CVD), diabetes, and the number of comorbidities, in the two datasets where the relevant information was available (InCHIANTI and WHAS). We performed analyses with DM calculated using either DM9, DM17, or the full set (31 biomarkers; DM31), allostatic load, and SAH. Allostatic load was calculated as closely as possible to previous publications [42,43] with the available biomarkers in InCHIANTI and WHAS; see Additional File 1 for details. To make the scales of DM, allostatic load, and SAH more comparable, we divided each score by its standard deviation (SD). Due to data availability differences, frailty was analyzed longitudinally in WHAS and cross-sectionally in InCHIANTI, while chronic diseases (CVD and diabetes) and the number of comorbidities were analyzed cross-sectionally for WHAS but longitudinally for InCHIANTI (see Additional File 1 for details). The relationship between dysregulation scores and mortality was assessed using time-to-event Cox proportional hazards models with age as the timescale. To study the relationship with frailty criteria and number of comorbidities we used Poisson regression. Logistic regression was used for individual chronic diseases. Age was rigorously controlled for using

a flexible cubic basis spline (`bs` function, `fda` package) with five degrees of freedom in InCHIANTI and four in WHAS. Poisson regressions and logistic regressions were implemented with Bayesian mixed models (`MCMCglmm` package) when longitudinal data were available, controlling for individual as a random effect (see Additional File 1 for details). For cross-sectional data, we used the `glm` function. We tested the relationship between dysregulation scores and each health outcome with three different models: 1) models that controlled for age and sex; 2) models that controlled for age, sex, and socio-economic status (education level in InCHIANTI; race, income, and education level in WHAS); and 3) models that controlled for age, sex, and metrics of physical and cognitive functions (Mini Mental State Examination score and time to walk four meters). In all models, we used a composite RP composed of an equal number of individuals from InCHIANTI, BLSA, and NHANES (3414 subjects in total) for DM calculation (see Additional File 1 for details). For InCHIANTI, SAH ranged from "very poor" to "very good", whereas in WHAS, it ranged from "excellent" to "poor", both on a scale from 1 to 5. To facilitate comparison, we used inverted SAH scores for InCHIANTI, such that lower scores indicate better health and higher scores worst health, mirroring DM. Analyses were performed in R-3.2.2 and codes are available upon request.

Results

Establishment of biomarker suites

We followed a detailed procedure (see Supplementary Methods section 1.5, Additional File 1) to choose subsets of biomarkers that would provide a more stable version of DM. In particular, we first narrowed the list of 31 biomarkers down to 22 by eliminating those with

means that varied greatly across datasets relative to variability. Among these 22, we tested each combination of 5 or 10 biomarkers, and then evaluated the impact of including/excluding a biomarker in a combination on (a) how robust DM was to choice of reference population, and (b) how closely correlated it was with the full 31-biomarker version (Supplementary Table 2 and Supplementary Fig. 2, Additional File 1). For example, folate introduced a strong dependency of the signal on the reference population, probably due to fortification policies in the U.S., and was thus not retained in the final list. Other subject criteria (data availability, breadth of physiological representation) were also considered to arrive at final lists of 9 and 17 markers.

Stability of DM9 and DM17 according to the choice of the reference population

DM calculated with both final sets (DM9 and DM17) proved to be highly stable, i.e. the signal did not vary substantially across various definitions of the RP (Fig. 1 and Supplementary Figs. 3-5, Additional File 1). Figure 1 shows that DM9 and DM17 are more stable than DM31 when calculated using other datasets as the RP, with respective mean correlation coefficients of 0.95, 0.95, and 0.86. These results show that by restricting ourselves to biomarkers that vary less across different populations, we obtained a stable signal regardless of the choice of the reference population. DM calculated using various demographic subsets of the study population as the RP is similarly stable (mean correlation coefficients of 0.96, 0.96, and 0.95, respectively for DM9, DM17, and DM31, see Supplementary Figs. 3-5, Additional File 1).

Association of DM9 and DM17 with health outcomes

Figs 2 and 3 show the associations of the various health metrics with health outcomes in InCHIANTI and WHAS, respectively. Generally speaking, all five metrics (DM9, DM17, DM31, allostatic load, and SAH) are competitive in their predictive ability, with some performing better in one analysis than another, but no clear "winner." DM31 generally performed a bit better than DM17, which performed a bit better than DM9, as expected. All metrics are comparable for mortality prediction (mean hazard ratios of 1.21, 1.29, 1.27, 1.24, and 1.33 respectively for DM9, DM17, DM31, AL, and SAH); however, DM-based metrics tend to show less variation across datasets (Fig. 4). SAH appears to be more strongly associated with frailty than the biomarker-based metrics: estimated regression coefficients were of 0.41 and 0.29, respectively for InCHIANTI and WHAS, whereas other metrics only reached ~0.15 (Fig.4). In InCHIANTI, DM31 appeared to perform particularly well for CVD and diabetes prediction (mean odds ratios of 1.7 and 2.6, respectively), likely reflecting the inclusion of metabolic-syndrome-related biomarkers in this version. Similarly, the high performance of AL for diabetes prediction (mean odds ratio of 2.0) might be due to the inclusion of glucose in its calculation, as opposed to DM9 and DM17. Prediction of the number of comorbidities is also relatively similar across metrics, with mean estimated regression coefficients of 0.05, 0.10, 0.17, 0.11, and 0.24, respectively for DM9, DM17, DM31, AL, and SAH (Fig. 4).

Discussion

We have previously proposed a metric of physiological dysregulation (DM), based on statistical distance and relying exclusively on common clinical biomarkers [13]. Here we aimed to reduce the number of biomarkers used in its calculation so that DM can be used in

contexts where fewer biomarkers are available (e.g. in socio-economic studies) and to propose a version of DM that is highly stable across different populations, so that it can be easily compared across studies. We had previously shown that DM's signal increases with the number of biomarkers included, although the value of additional markers diminishes as more are added [17], and that inclusion of 10-15 is generally sufficient. Using solely biomarkers from the complete blood count, the lipid and liver panels, as well as calcium and vitamin B12, we identified and validated two DM versions: a version using 17 biomarkers and a shorter version that uses only 9 biomarkers, excluding the ones that may be slightly less common (GGT, ALKP, AST, albumin, total proteins, calcium, and vitamin B12). Nine or 17 markers may seem like a lot, but eight are measured together in the complete blood count, while HDL is highly common, and many of the liver proteins are measured together in a panel; many existing studies already have all these markers. Both versions proved to be highly stable across various definitions of the RP and to provide good predictions of health outcomes, though the 17-biomarker version performs slightly better for prediction. We thus propose these dysregulation signatures ("DSign") as generalized, objective metrics of health state, with DM17 to be preferred when possible.

As expected, there was no clear "winner" among metrics of health state to predict various health outcomes. Some metrics performed better for certain outcomes or in one or the other dataset. For example, SAH performs best for predicting phenotypic frailty, an unsurprising result given that phenotypic frailty is diagnosed based on physical symptoms a patient would recognize rather than on measurement of the underlying pathology. Likewise, as expected, DM31 generally performs as well as or better than DM17 and DM9 for predicting health outcomes, particularly diabetes and CVD, which are related to some of the metabolic-

syndrome-associated biomarkers that were eliminated in order to increase robustness of the signal. Nonetheless, for mortality, all metrics perform about equivalently. Interestingly, in most cases the strengths of the effects were minimally impacted by control for covariates, including socioeconomic status and markers of physical and cognitive functioning. This was true not just for versions of DM, but also allostatic load and SAH. Potentially, this is due the underlying health state mediating the impacts of the covariates on the outcomes.

Health plays an important role in many study fields and efforts have been made in the search for robust and comparable health metrics. While many existing health metrics are good predictors of mortality, frailty and comorbidities, notably SAH, we believe it is meaningful to have an objective and continuously distributed metric of general health based on continuous variables (biomarkers). First, a continuous health metric can facilitate the estimation of the distribution of health states. Indices of health inequality can also be easily calculated with the continuous health metric. For example, the concentration index has become a standard metric to quantify income-related inequalities [44]. Strictly speaking, the concentration index is an appropriate metric of socioeconomic-related health inequality when health is measured on a ratio scale with a true zero [45]. Our health metric satisfies these requirements by definition, where a value of zero represents the ideal state of health. An application to the calculation of the concentration index was illustrated in a working paper [46]. Second, a continuous health metric facilitates the use of certain statistical tools, such as ordinary least squares or instrumental variable regression, whose consistency relies less on distributional assumptions [47]. Single biomarkers have occasionally been used as indicators of health outcome in statistical models that require a continuous health variable [48]; however, it would be preferable to summarize the information from multiple

biomarkers into a single metric when measuring global health. Third, in comparison with subjective health metrics (e.g. self-reported health) or quasi-objective health metrics (e.g. composite health metrics constructed from survey questions) the health metric here could be applied more easily across different populations without being influenced by cultural differences or reporting habits. Indeed, several studies have reported differences in rating health according to gender [49,50], ethnicity [51,52], and age [53,54]. Last but not least, in keeping the biomarkers continuous during the construction of the health metric, we may well avoid loss of information associated with categorization of continuous variables [55]. In cases where the study population is small or broadly representative of the population, we strongly recommend using our reference population; however, in cases where the study population is both large enough to serve as its own reference, and is highly specific (e.g. suffering from a particular disease, children, a non-industrialized tribe), we would recommend using the study population as the reference population.

It is important to note several limitations of this approach as well. First, we would not recommend application of this DSign metric to populations suffering from a specific disease. For example, a study on the efficacy or safety of a medication for patients on hemodialysis should not rely on DM17 or DM9 as a proxy outcome, because hemodialysis *a priori* represents a state of extreme dysregulation of multiple biomarkers [56], for which our standardized RP would be inappropriate without independent validation. Second, there are clearly multiple dimensions to physiological health, and any single metric is by definition a crude simplification [30,57]. The advantages of this approach should not be used to gloss over the limitations of any such approach. Third, the advantages of this approach do not make it the best choice in all cases. For example, SAH may be a preferable representation of

394 health state in some cases, either for practical reasons (e.g. better prediction of frailty,
395 empirically) or theoretical reasons (e.g. a specific interest in how perception of one's health
396 influences outcomes). Fourth, we do not claim that the version presented here is the only
397 valid version of DM, or necessarily the best; it is one approach among many that appears to
398 represent a nearly optimal balance of usability, stability, and predictive value, but
399 sophisticated users may prefer to develop their own versions based on data availability or
400 their specific needs for these sometimes conflicting factors. Fifth, the populations used to
401 establish stability here, while from two continents, nonetheless represent modern, Western
402 societies. Caution should be exercised applying the metric to other populations, though
403 studies have shown that DM does work well as a health metric in several provinces in
404 Chinese mainland, in Taiwan, and in the Tsimane horticulturalists of Bolivia [18–20,27]. We
405 believe it would probably apply well in most contexts, but maybe not in populations with
406 highly specific characteristics (e.g. non-industrial populations [58]).
407
408 **Conclusions**
409 We have developed a continuous, biomarker-based, standardized, validated metric of health
410 state. While no single metric can be universally optimal, this metric presents a number of
411 clear advantages: simplicity of use, ease to obtain the relevant biomarkers, predictive power
412 competitive with other well-known metrics, stability across populations, and theoretical non-
413 circularity. For many users, it will present a substantial improvement over previously
414 published versions of DM, notably in its standardization and stability. We nonetheless
415 strongly urge users of any generalized health metric to use caution and a nuanced

interpretation, given the inherent challenges of using a single metric to measure a multi-dimensional process in a complex system.

Abbreviations

AL: allostatic load

ALKP: alkaline phosphatase

ALT: alanine transaminase

AST: aspartate transaminase

BASO%: basophil percentage

BLSA: Baltimore Longitudinal Study of Aging

DM: Mahalanobis distance (dysregulation score)

DM9: 9-set DM

DM17: 17-set DM

DM31: 31-set DM

GGT: gamma-glutamyl transferase

HDL: high density lipoprotein

InCHIANTI: Invecchiare in Chianti

LDH: lactate dehydrogenase

LYM%: lymphocyte percentage

MCH: mean corpuscular hemoglobin

MCHC: mean corpuscular hemoglobin concentration

MONO%: monocyte percentage

NEUT%: neutrophil percentage

NHANES: National Health and Nutrition Examination Survey

RBC: red blood cell count

RDW: red cell distribution width

RP: Reference population

SAH: self-assessed health

SD: standard deviation

WBC: white blood cells

WHAS: Women's Health and Aging Study

Declarations

Ethics approval and consent to participate

All aspects of dataset collection were approved by the ethics committees at the institutions responsible for data collection, or by the National Institute of Environmental Health Services Internal Review Board for BLSA, and this secondary analysis was approved by the ethics committee (*Comité d'éthique de la recherche en santé chez l'humain*) at the *Centre de recherche du CHUS*, project # 14-059. Participants signed informed consent for data collection and analysis.

Consent for publication

Not applicable.

Availability of data and material

With the exception of NHANES, the data used in these analyses cannot be freely shared due to confidentiality constraints related to human medical data, but they are all available to

researchers submitting an appropriate research proposal: InCHIANTI at http://www.inchiantistudy.net/obtain_data.html, WHAS at https://jhpeppercenter.jhmi.edu/ec_proposal/login.aspx, and BLSA at http://www.blsa.nih.gov/researchers. NHANES data is available at https://www.cdc.gov/nchs/nhanes/.

Competing interests

AAC declares a CoI as Founder and CEO at Oken Health. No other competing interests are declared.

Funding

This work was supported by the Canadian Institutes of Health Research (CIHR, grant numbers 119485 and 145585). QL is supported by the National Natural Science Foundation of China (82060265). AAC is supported by a Fonds de recherche du Québec – Santé (FRQ-S) Senior Salary Award and is a member of the Fonds de recherche du Québec – Santé funded Centre de recherche du CHUS and Centre de recherche sur le vieillissement.

Authors' contributions

LF collected the data for InCHIANTI and BLSA, and LPF for WHAS. QL and AAC conceived and designed the article. QL, VDG, and VL analyzed the data. AAC, QL and VL participated in data interpretation and writing of the manuscript. All authors have critically reviewed the manuscript, and they have read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Lindeboom M, van Doorslaer E. Cut-point shift and index shift in self-reported health. J Health Econ. 2004;23:1083–99.

2. Bound J. Self-reported versus objective measures of health in retirement models. Journal of Human Resources. 1991;26:106–38.

3. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. Journal of Clinical Epidemiology. 1994;47:1245–51.

4. Fortin M, Bravo G, Hudon C, Lapointe L, Almirall J, Dubois M-F, et al. Relationship Between Multimorbidity and Health-Related Quality of Life of Patients in Primary Care. Quality of Life Research. 2006;15:83–91.

5. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci. 2001/03/17. 2001;56:M146-56.

6. Mitnitski AB, Mogilner AJ, Rockwood K. Accumulation of deficits as a proxy measure of aging. ScientificWorldJournal. Hindawi Publishing Corporation; 2001;1:323–36.

7. Rockwood K, Song X, MacKnight C, Bergman H, Hogan DB, McDowell I, et al. A global clinical measure of fitness and frailty in elderly people. CMAJ. 2005;173:489–95.

8. Panhwar YN, Naghdy F, Naghdy G, Stirling D, Potter J. Assessment of frailty: a survey of quantitative and clinical methods. BMC Biomedical Engineering. 2019;1:7.

9. McEwen BS. Stress, adaptation, and disease. Allostasis and allostatic load. Ann N Y Acad Sci. 1998;840:33–44.

511 10. Seeman TE, McEwen BS, Rowe JW, Singer BH. Allostatic load as a marker of
512 cumulative biological risk: MacArthur studies of successful aging. Proc Natl Acad Sci U S
513 A. 2001;98:4770–5.
514 11. Seplaki CL, Goldman N, Glei D, Weinstein M. A comparative analysis of measurement
515 approaches for physiological dysregulation in an older population. Exp Gerontol.
516 2005/05/28. 2005;40:438–49.
517 12. Singer BH, Ryff CD, Seeman T. Operationalizing allostatic load. In: Schuli J, editor.
518 Allostasis, Homeostasis, and the Costs of Physiological Adaptation. New York, NY:
519 Cambridge University Press; 2004. p. 113–49.
520 13. Cohen AA, Milot E, Yong J, Seplaki CL, Fülöp T, Bandeen-Roche K, et al. A novel
521 statistical approach shows evidence for multi-system physiological dysregulation during
522 aging. Mech Ageing Dev. 2013/02/05. 2013;134:110–7.
523 14. Cohen AA, Milot E, Li Q, Legault V, Fried LP, Ferrucci L. Cross-population validation
524 of statistical distance as a measure of physiological dysregulation during aging. Exp
525 Gerontol. 2014/05/08. 2014;57:203–10.
526 15. Milot E, Morissette-Thomas V, Li Q, Fried LP, Ferrucci L, Cohen AA. Trajectories of
527 physiological dysregulation predicts mortality and health outcomes in a consistent manner
528 across three populations. Mech Ageing Dev [Internet]. 2014/12/03. 2014;141–142:56–63.
529 Available from: http://www.ncbi.nlm.nih.gov/pubmed/25454986
530 16. Cohen AA. Complex systems dynamics in aging: new evidence, continuing questions.
531 Biogerontology [Internet]. 2016;17:205–20. Available from:
532 https://www.ncbi.nlm.nih.gov/pubmed/25991473

17. Cohen AA, Li Q, Milot E, Leroux M, Faucher S, Morissette-Thomas V, et al. Statistical distance as a measure of physiological dysregulation is largely robust to variation in its biomarker composition. PLoS ONE. Public Library of Science; 2015;10.

18. Kraft TS, Stieglitz J, Trumble BC, Garcia AR, Kaplan H, Gurven M. Multi-system physiological dysregulation and ageing in a subsistence population. Philosophical Transactions of the Royal Society B: Biological Sciences [Internet]. Royal Society; 2020;375:20190610. Available from: https://doi.org/10.1098/rstb.2019.0610

19. Nie P, Li Q, Cohen AA, Sousa-Poza A. In search of China's income-health gradient: a biomarker-based analysis. Applied Economics [Internet]. Routledge; 2021;1–20. Available from: https://doi.org/10.1080/00036846.2021.1927962

20. Liu Z. Development and validation of two composite aging measures using routine clinical biomarkers in the Chinese population: Analyses from two prospective cohort studies. The Journals of Gerontology: Series A. 2020;

21. Arbeev KG, Cohen AA, Arbeeva LS, Milot E, Stallard E, Kulminski AM, et al. Optimal Versus Realized Trajectories of Physiological Dysregulation in Aging and Their Relation to Sex-Specific Mortality Risk. Front Public Health [Internet]. 2016/02/03. 2016;4:3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26835445

22. Belsky DWW, Moffitt TEE, Cohen AAA, Corcoran DLL, Levine MEE, Prinz JAA, et al. Eleven Telomere, Epigenetic Clock, and Biomarker-Composite Quantifications of Biological Aging: Do They Measure the Same Thing? Am J Epidemiol [Internet]. 2018;187:1220–30. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29149257

23. Belsky DW, Huffman KM, Pieper CF, Shalev I, Kraus WE, Anderson R. Change in the rate of biological aging in response to caloric restriction: Calerie Biobank analysis. Journals of Gerontology - Series A Biological Sciences and Medical Sciences [Internet]. Oxford

University Press; 2018 [cited 2020 Dec 2];73:4–10. Available from: https://pubmed.ncbi.nlm.nih.gov/28531269/

24. Parker DC, Bartlett BN, Cohen HJ, Fillenbaum G, Huebner JL, Kraus VB, et al. Association of Blood Chemistry Quantifications of Biological Aging With Disability and Mortality in Older Adults. The Journals of Gerontology: Series A [Internet]. 2020;75:1671–9. Available from: https://doi.org/10.1093/gerona/glz219

25. Ghachem A, Fried LP, Legault V, Bandeen-Roche K, Presse N, Gaudreau P, et al. Evidence from two cohorts for the frailty syndrome as an emergent state of parallel dysregulation in multiple physiological systems. Biogerontology [Internet]. Springer Science and Business Media B.V.; 2021 [cited 2021 Mar 21];22:63–79. Available from: https://link.springer.com/article/10.1007/s10522-020-09903-w

26. Arbeev KG, Bagley O, Ukraintseva S v, Duan H, Kulminski AM, Stallard E, et al. Composite Measure of Physiological Dysregulation as a Predictor of Mortality: The Long Life Family Study [Internet]. Frontiers in Public Health . 2020. p. 56. Available from: https://www.frontiersin.org/article/10.3389/fpubh.2020.00056

27. Gaydosh L, Belsky DW, Glei DA, Goldman N. Testing Proposed Quantifications of Biological Aging in Taiwanese Older Adults. The Journals of Gerontology: Series A [Internet]. 2020;75:1680–5. Available from: https://doi.org/10.1093/gerona/glz223

28. Dansereau G, Wey TW, Legault V, Brunet MA, Kemnitz JW, Ferrucci L, et al. Conservation of physiological dysregulation signatures of aging across primates. Aging Cell. 2019;

29. Milot E;, Cohen AA;, Vézina F;, Buehler DM;, Matson KD;, Piersma T. A novel integrative method for measuring body condition in ecological studies based on physiological dysregulation. Methods in Ecology and Evolution. 2014;5:146–155.

30. Li Q, Wang S, Milot E, Bergeron P, Ferrucci L, Fried LP, et al. Homeostatic dysregulation proceeds in parallel in multiple physiological systems. Aging Cell. 2015/09/30. 2015;14:1103–12.

31. Weiss CO, Cappola AR, Varadhan R, Fried LP. Resting metabolic rate in old-old women with and without frailty: variability and estimation of energy requirements. Journal of the American Geriatrics Society. 2012;60:1695–700.

32. Kalyani RR, Varadhan R, Weiss CO, Fried LP, Cappola AR. Frailty status and altered glucose-insulin dynamics. J Gerontol A Biol Sci Med Sci. Oxford University Press; 2012;67:1300–6.

33. Shock NW. Normal Human Aging: The Baltimore Longitudinal Study of Aging. Vol NIH-84-2450 661 (National Institute of Aging, Washington DC). 1984;

34. Ferrucci L. The Baltimore Longitudinal Study on Aging: a 50 year long journey and plans for the future. Giornale di Gerontologia. 2009;57:3–8.

35. Ferrucci L, Bandinelli S, Benvenuti E, Di Iorio A, Macchi C, Harris TB, et al. Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. J Am Geriatr Soc. 2000/12/29. 2000;48:1618–25.

36. Centers for Disease Control and Prevention. National Healthand Nutrition Examination Survey: Plan and Operations, 1999–2010. U.S. Department of Health and Human Services; 2013.

37. Fried LP, Kasper KD, Guralnik JM, Simonsick EM. The Women's Health and Aging Study: an introduction. In: Guralnik JM, Fried LP, Simonsick EM, Kasper KD, Lafferty ME, editors. The Women's Health and Aging Study: health and social characteristics of old women with disability. Bethesda, MD: National Institute on Aging; 1995. p. 1-8.

605 38. Guralnik Fried L.P. Kasper J.D. Simonsick E.M. & Lafferty M.E. JM, Guralnik Fried,
606 L.P., Kasper, J.D., Simonsick, E.M. & Lafferty, M.E. JM. The Women's Health and Aging
607 Study: health and social characteristics of older women with disability. Washington DC:
608 National Institute on Aging; 1995.
609 39. Mahalanobis PC; Mahalanobis distance. Proc Natl Inst Sci India. 1936;49:234–256.
610 40. le Couteur DG, Simpson SJ. Adaptive senectitude: The prolongevity effects of aging.
611 Journals of Gerontology - Series A Biological Sciences and Medical Sciences. 2011.
612 41. Cohen AA, Levasseur M, Raina PS, Fried LP, Fulop T. Is aging biology ageist? Journals
613 of Gerontology - Series A Biological Sciences and Medical Sciences.
614 42. Juster RP, McEwen BS, Lupien SJ. Allostatic load biomarkers of chronic stress and
615 impact on health and cognition. Neurosci Biobehav Rev. 2010;35:2–16.
616 43. Sibille KT, McBeth J, Smith D, Wilkie R. Allostatic load and pain severity in older
617 adults: Results from the English Longitudinal Study of Ageing. Exp Gerontol. 2017;88:51–
618 8.
619 44. Wagstaff A;, Van Doorslaer E. Equity in health care finance and delivery. In: Culyer A;,
620 Newhouse J, editors. Handbook of Health Economics. Amsterdam: North Holland; 2000. p.
621 1804–62.
622 45. O'Donnell E.; Wagstaff A.; Lindelow M. O; van D, O'Donnell E.; Wagstaff, A.;
623 Lindelow, M. O; van D. Analyzing Health Equity Using Household Survey Data : A Guide
624 to Techniques and Their Implementation. Washington, DC: World Bank; 2008.
625 46. Li Q;, Cohen AA;, Legault V;, Girard VD;, Ferrucci L;, Fried LP. An objective measure
626 of individual health and aging for population surveys.
627 47. Wooldridge JM. Econometric analysis of cross section and panel data. 2nd ed.
628 Cambridge, Mass.: MIT Press; 2010.

629 48. Li Q. Identifiability of mean-reverting measurement error with instrumental variable.
630 Statistica Neerlandica. 2014;68:118–29.
631 49. Benyamini Y, Leventhal EA, Leventhal H. Gender Differences in Processing
632 Information for Making Self-Assessments of Health. Psychosomatic Medicine. 2000;62.
633 50. Crimmins EM, Kim JK, Solé-Auró A. Gender differences in health: results from
634 SHARE, ELSA and HRS. European Journal of Public Health. 2011;21:81–91.
635 51. Alang SM, McCreedy EM, McAlpine DD. Race, Ethnicity, and Self-Rated Health
636 Among Immigrants in the United States. Journal of racial and ethnic health disparities.
637 Switzerland; 2015;2:565–72.
638 52. Zhang H, Bago d'Uva T, van Doorslaer E. The gender health gap in China: A
639 decomposition analysis. Economics & Human Biology. 2015;18:13–26.
640 53. Liang J, Shaw BA, Krause N, Bennett JM, Kobayashi E, Fukaya T, et al. How Does
641 Self-Assessed Health Change With Age? A Study of Older Adults in Japan. The Journals of
642 Gerontology: Series B. 2005;60:S224–32.
643 54. Pinquart M. Correlates of subjective health in older adults: A meta-analysis. Psychology
644 and Aging. US: American Psychological Association; 2001. p. 414–26.
645 55. Barnwell-Menard JL, Li Q, Cohen AA. Effects of categorization method, regression
646 type, and variable distribution on the inflation of Type-I error rate when categorizing a
647 confounding variable. Stat Med. 2014/12/17. 2015;34:936–49.
648 56. Nakazato Y, Kurane R, Hirose S, Watanabe A, Shimoyama H. Aging and death-
649 associated changes in serum albumin variability over the course of chronic hemodialysis
650 treatment. PLOS One [Internet]. 2017;12:e0185216. Available from:
651 https://www.ncbi.nlm.nih.gov/pubmed/28953942

57. Fried LP, Xue Q-LL, Cappola AR, Ferrucci L, Chaves P, Varadhan R, et al. Nonlinear multisystem physiological dysregulation associated with frailty in older women: implications for etiology and treatment. J Gerontol A Biol Sci Med Sci. Oxford University Press; 2009;64:1049–57.

58. Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol. 2016;17:171.

Additional files

Additional file 1. Supplementary methods, tables and figures. (DOCX)

Figure legends

Figure 1. Stability of dysregulation scores across populations. For each dataset or a combined set (All), we performed correlations between dysregulation scores (DM) calculated using the study population its own reference population (top) or another dataset as the reference population (left). Correlations were calculated for the three biomarker sets: 9 biomarker-set (DM9), 17-set (DM17), and the entire set (DM31). Mean correlation coefficients (r) are indicated for each set and ellipses indicate correlations visually, i.e. darker and narrower when stronger.

Figure 2. Relationships between health metrics and aging correlates in the InCHIANTI dataset. Estimations (points) together with 95% confidence intervals (CIs; segments) are plotted for mortality, the number of frailty criteria, cardiovascular diseases (CVD), diabetes,

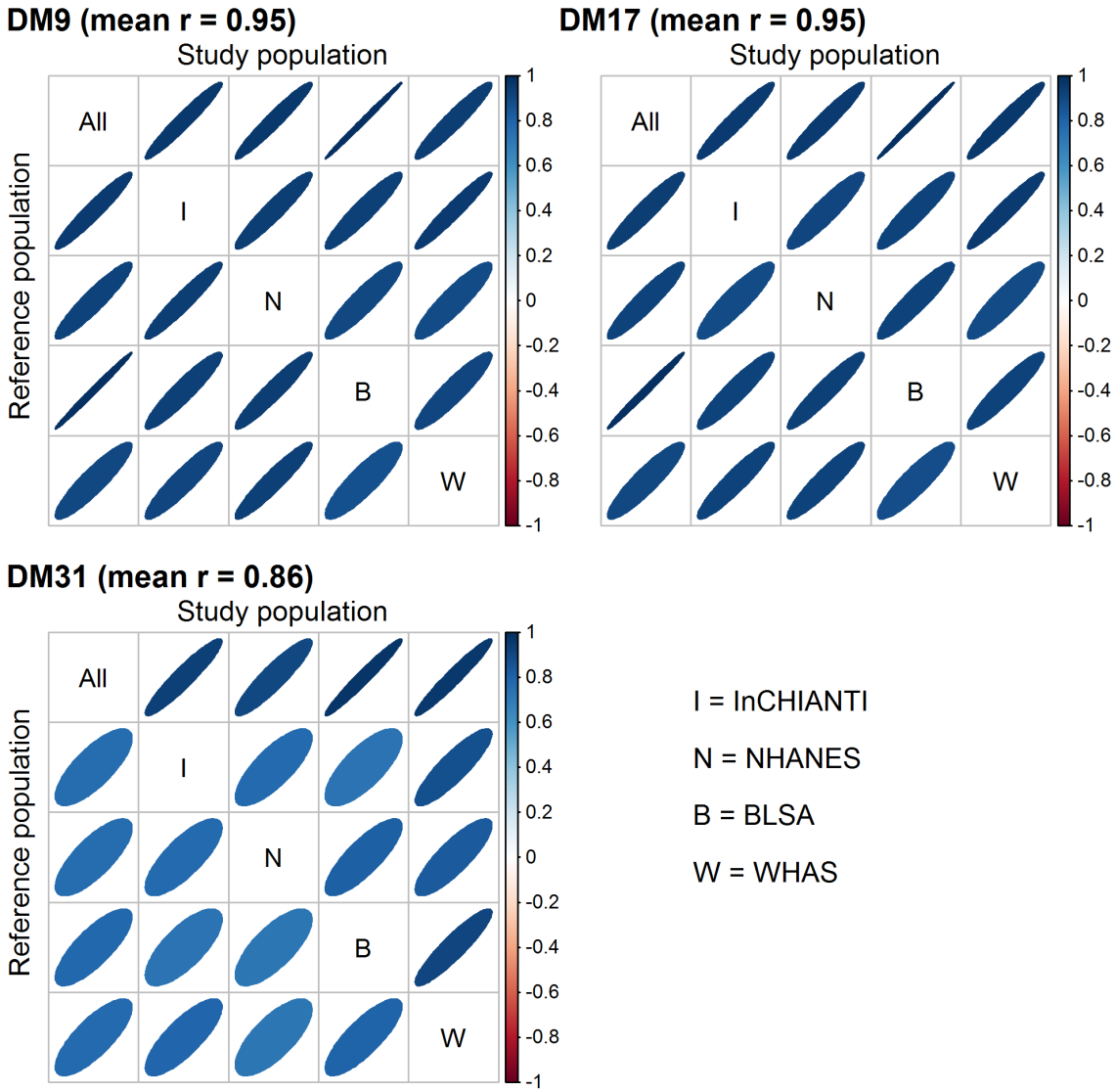
676 and the number of comorbidities (see text for details). Results are based on regression models
677 adjusting for: 1) age and sex (solid lines); 2) age, sex, as well as physiological and cognitive
678 functions (dashed lines); or age, sex, and socioeconomic status (dotted lines). For ease of
679 comparison, each metric was standardized, i.e. divided by its standard deviation. Different
680 colors refer to different health metrics and estimates are indicated on the right. Significant
681 results are plotted in bold, with asterisks indicating the significance level (***, $p < 0.001$; **,
682 $p < 0.01$; *, $p < 0.05$). Abbreviations: AL, allostatic load; DM9, 9-set dysregulation score
683 (DM); DM17, 17-set DM; DM31, 31-set DM; SAH, self-assessed health.
684
685 **Figure 3. Relationships between health metrics and aging correlates in the WHAS**
686 **dataset.** Estimations (points) together with 95% confidence intervals (CIs; segments) are
687 plotted for mortality, the number of frailty criteria, cardiovascular diseases (CVD), diabetes,
688 and the number of comorbidities (see text for details). Results are based on regression models
689 adjusting for: 1) age (solid lines); 2) age as well as physiological and cognitive functions
690 (dashed lines); or age and socioeconomic status (dotted lines). For ease of comparison, each
691 metric was standardized, i.e. divided by its standard deviation. Different colors refer to
692 different health metrics and estimates are indicated on the right. Significant results are plotted
693 in bold, with asterisks indicating the significance level (***, $p < 0.001$; **, $p < 0.01$; *, $p <$
694 0.05). Abbreviations: AL, allostatic load; DM9, 9-set dysregulation score (DM); DM17, 17-
695 set DM; DM31, 31-set DM; SAH, self-assessed health.
696
697
698 **Figure 4. Comparison of predictive performance across health metrics for various health**
699 **outcomes.** Bars represent the means of estimated regression coefficients for the three different

700 analyses performed (see Figs 2-3) in InCHIANTI (blue) and WHAS (red), with the
701 corresponding 95% confidence interval. For ease of comparison across health outcomes, we
702 used the log-hazard and log-odds ratios. Numbers above the bars indicate the number of
703 significant associations out of three analyses. Abbreviations: Comorb., number of
704 comorbidities; DM9, 9-set dysregulation score (DM); DM17, 17-set DM; DM31, 31-set DM.

Tables

Table 1. Characteristics of study populations (at first visit).

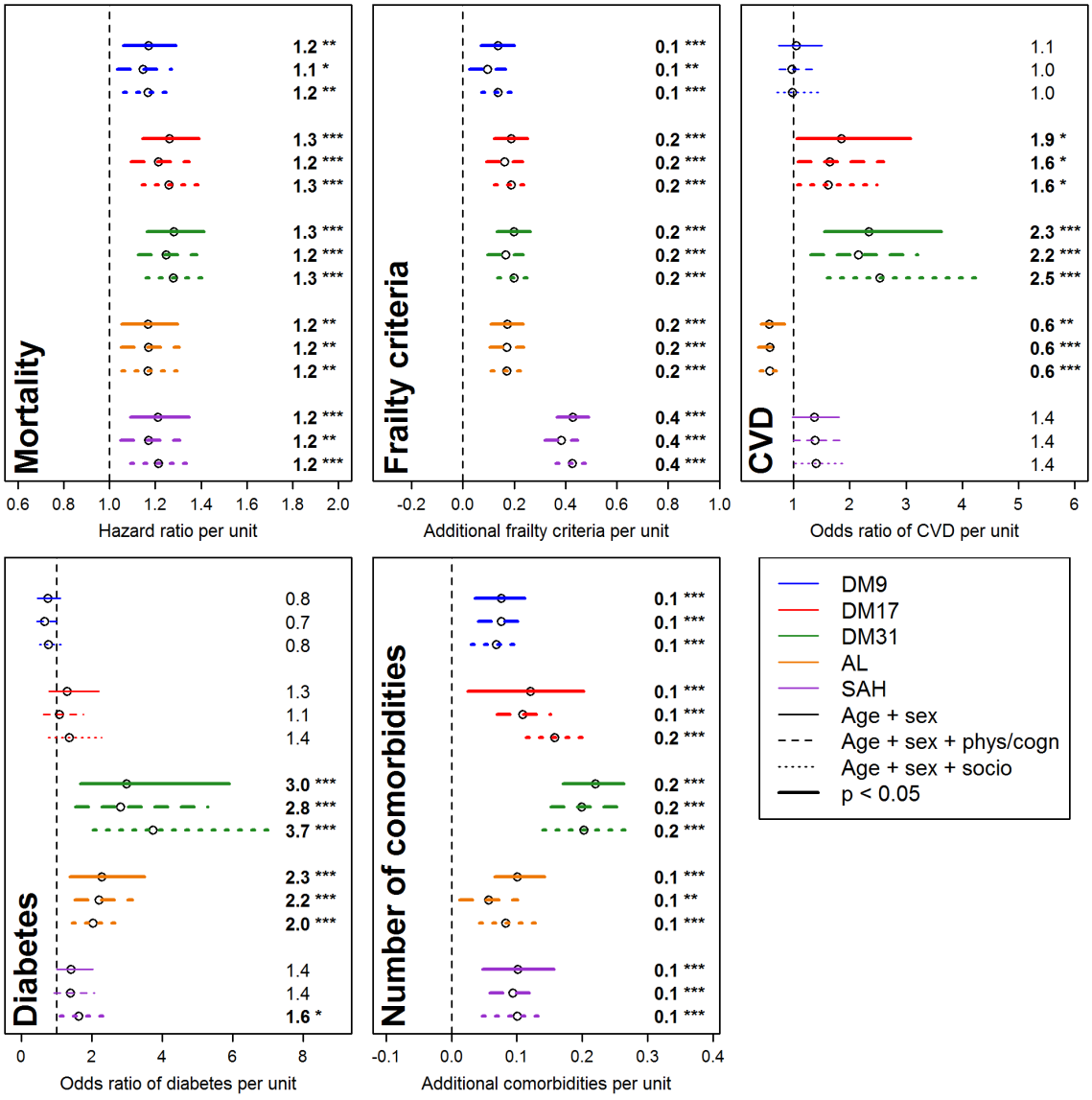
Characteristic	BLSA n = 1139	InCHIANTI n = 1252	NHANES n = 17,379	WHAS n = 1067
Age (years)				
Mean ± SD	64.6 ± 13.8	68.2 ± 15.5	49.4 ± 19.0	77.1 ± 6.8
Range (min – max)	26.4 – 99.3	21.3 – 98.4	20 – 85	65.8 – 100.3
Female (%)	549 (48.2)	694 (55.4)	9073 (52.2)	1067 (100.0)
Race (white, %)	728 (63.9)	1252 (100.0)	8768 (50.5)	801 (75.1)
Education (years), mean ± SD	17.0 (2.6)	7.2 (14.5)	—	10.7 (3.8)
4-meter walking time (sec), mean ± SD	—	4.1 ± 2.8	—	9.8 ± 10.0
MMSE score, mean ± SD	—	25.9 ± 3.7	—	26.5 ± 3.0
Self-assessed health				
1 (%) – highest perceived health	—	159 (13.3)	—	25 (4.2)
2 (%)	—	640 (53.7)	—	88 (14.9)
3 (%)	—	323 (27.1)	—	188 (31.9)
4 (%)	—	57 (4.8)	—	199 (33.7)
5 (%) – lowest perceived health	—	13 (1.1)	—	90 (15.3)
Allostatic load				
Mean ± SD	—	2.6 ± 1.8	—	2.1 ± 1.6
Range (min – max)	—	0 – 11	—	0 – 9



708

709 **Figure 1.**

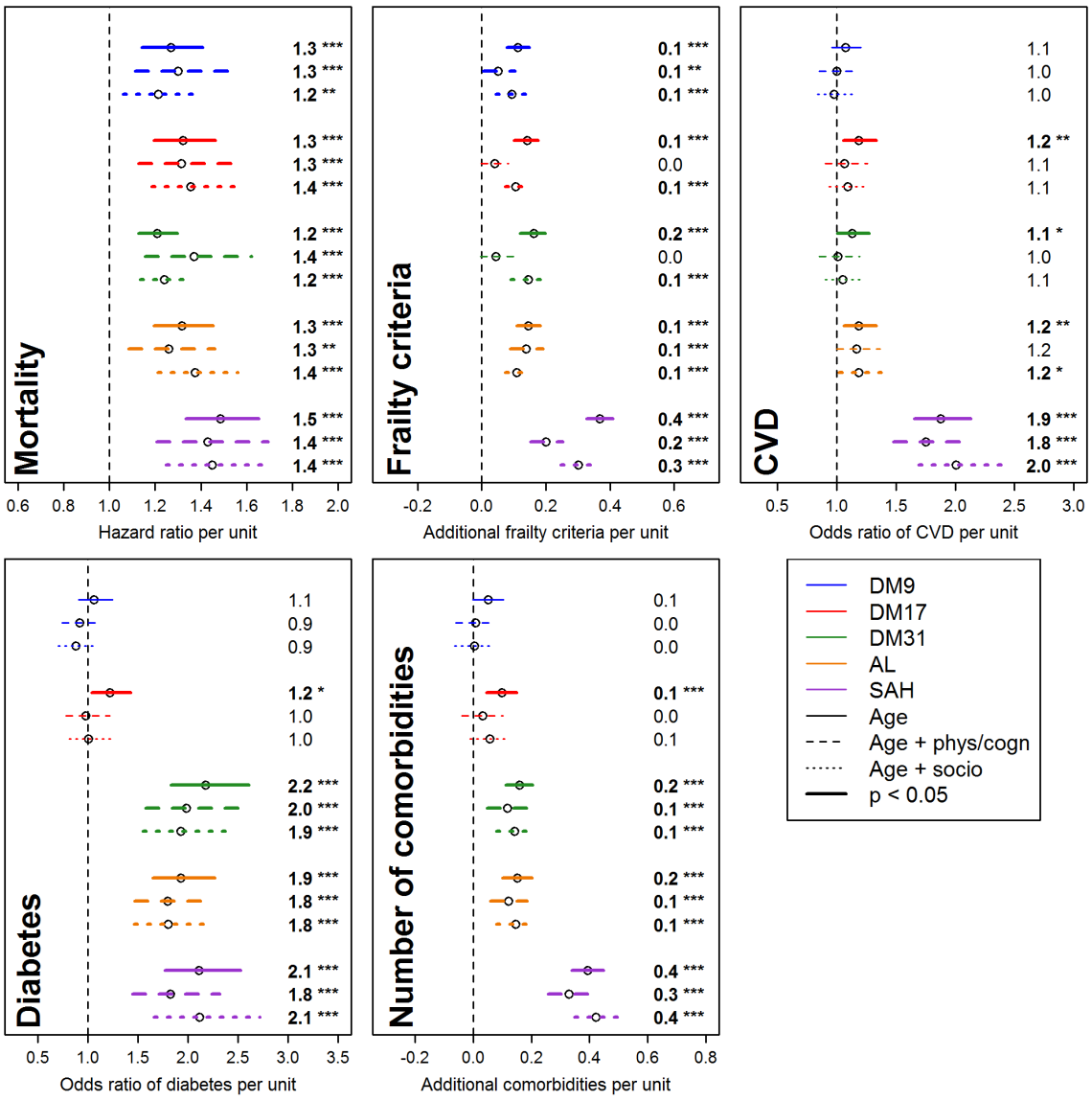
710



711

712 **Figure 2.**

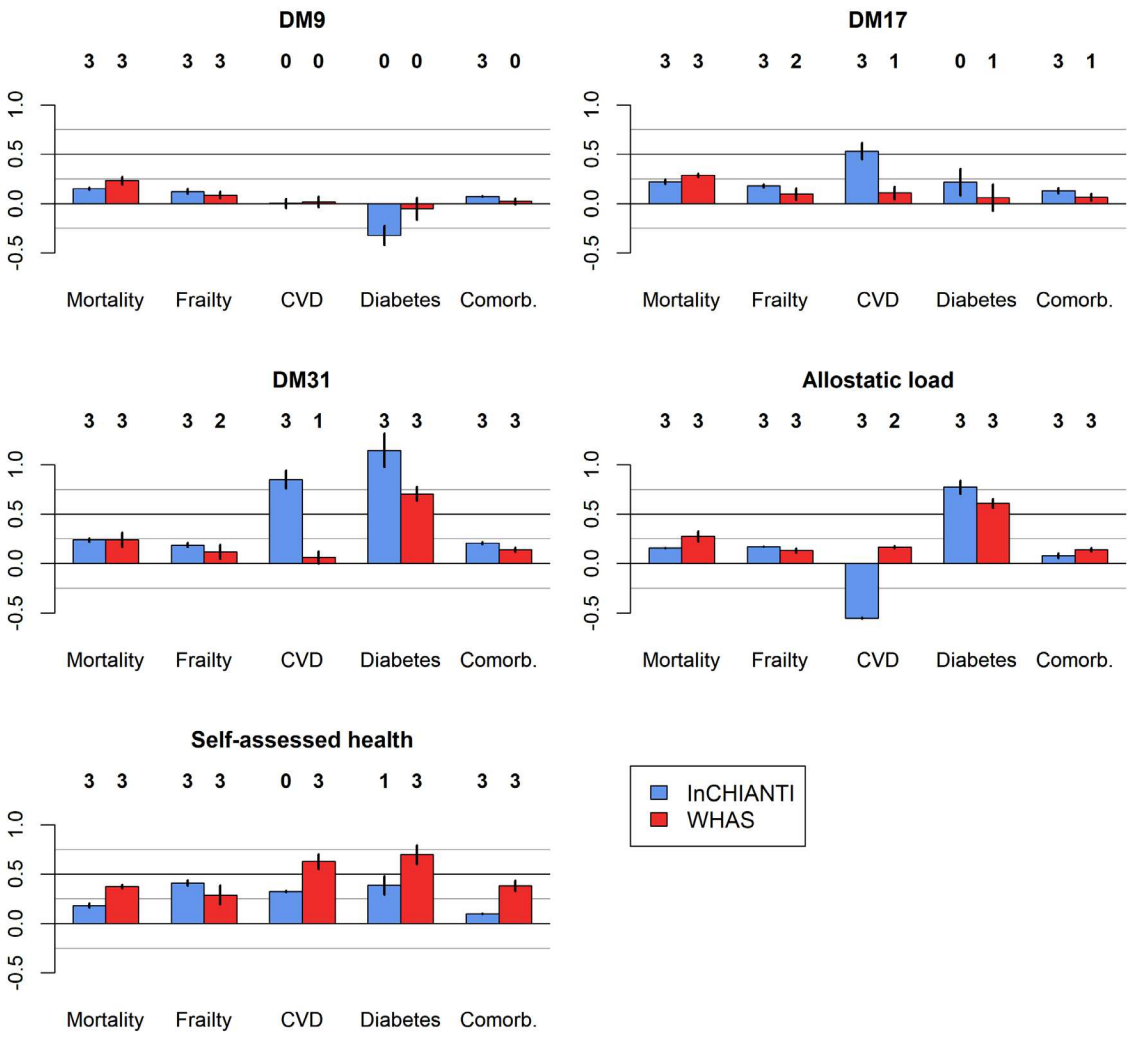
713



714

715 **Figure 3.**

716



717

718 **Figure 4.**

719

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PHMSIfinal.pdf