

Office of Population Research Princeton University
WORKING PAPER SERIES

A Comparative Analysis of Measurement Approaches for
Physiological Dysregulation in an Older Population

Working Paper No. 2004-04

Christopher L. Seplaki
Office of Population Research

Noreen Goldman
Office of Population Research

Dana Gleib
Georgetown University, Washington, D.C.

Maxine Weinstein
Georgetown University, Washington, D.C.

Correspondence to: Christopher L. Seplaki, Ph.D., cseplaki@princeton.edu, Office of Population Research, 263 Wallace Hall, Princeton University, Princeton, NJ 08544. Voice: 609.258.1394, Fax: 609.258.1039.

Short running page headline: Comparison of Allostatic Load Measures

Key Words: Aging; Allostatic load; Biomarkers; Health outcomes; Physical mobility; Cognitive function; Depression; Self-assessed health

Papers published in the OPR Working Paper Series reflect the view of individual authors. They may not be cited in other publications, but are intended to be work-in-progress. Comments are welcome. Additional copies are available by writing to the Office of Population Research Working Papers Series, Princeton University, Wallace Hall, Second floor, Princeton, NJ 08544 Fax (609) 258- 1039, or from the web, <http://www.opr.princeton.edu/>

Abstract

A growing body of evidence suggests that the cumulative experience of emotional challenges and stressful events over the lifecourse may take a significant physiological toll, referred to as allostatic load, on multiple interrelated systems of the body. Multiple summary measures of these effects have been proposed in the literature, but there is no systematic evaluation of alternative measurements. We use data from a population-based sample of older Taiwanese to compare the explanatory power and predictive performance of several measures of allostatic load for diverse health outcomes. We find that, although modest, the various assumptions made when constructing measures of allostatic load do have effects that should be considered carefully. Our findings suggest development of measures that preserve the continuous properties of the component biological measurements and underscore the importance of nonlinear effects suggested in previous research. These fundamental insights are of use to applied researchers in the field currently in search of useful empirical formulations of allostatic load and to those who are focused on the development of improved measurement strategies.

Abstract word count = 171

1 Introduction

A growing body of evidence suggests that the cumulative experience of emotional challenges and stressful events over the lifecourse may take a significant physiological toll on multiple interrelated systems of the body. The manifestation of this cumulative damage, referred to as *allostatic load* (McEwen, Stellar, 1993), is the disruption of otherwise balanced interactions among these diverse systems reflected by key biological variables that are outside optimal operating ranges. Just how the physiological dysregulation characterizing allostatic load should be measured using biological variables in population-based research remains a subject of debate. Here, we compare the explanatory power and predictive performance of several formulations to provide a critical point of reference for future development.

The system parameters that have been hypothesized to manifest allostatic load can be divided into groups that approximate a theoretical sequence of physiological events tying chronic stimulation of the stress response with disease endpoints (McEwen, 2002; McEwen, Seeman, 1999). In this framework hormonal factors are grouped together and defined as “primary mediators”. This group includes markers of stress-related sympathetic nervous system (SNS) activity (e.g., epinephrine and norepinephrine), hypothalamic-pituitary-adrenal (HPA) axis activity (e.g., cortisol), and related inflammation (e.g., interleukin-6) and growth hormone (e.g., insulin-like growth factor-1) responses. Additionally, researchers hypothesize that dehydroepiandrosterone sulfate (DHEA-S) plays an important mediating role among these factors; low levels are associated with both worse mental health and poorer physical functioning (Berr, Lafont, Debuire, Dartigues, Baulieu, 1996; Mazat, Lafont, Berr, Debuire, Tessier, Dartigues, Baulieu, 2001; Svec, Lopez, 1989). The framework postulates that these hormonal factors have “primary” effects on tissues and organs that lead to “secondary outcomes” at the

system level. These include metabolic syndrome (Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, 2001), which can be assessed by several cardiovascular disease risk factors (e.g., blood pressure, sugar, and cholesterol levels, body mass index, and waist-to-hip ratio). “Tertiary outcomes” are the disease endpoints that result from secondary outcomes (e.g., coronary heart disease).

Several summary measures of such multisystem physiological dysregulation associated with allostatic load have been proposed in the literature. The approach proposed first, and the one most commonly applied, is constructed as the sum of the number of biological measurements (out of a given set, most commonly ten) for which an individual falls into the highest risk quartile of each measurement’s distribution (Seeman, Singer, Rowe, Horwitz, McEwen, 1997). Researchers have demonstrated that this simple “count-based” formulation is associated with diverse health outcomes, including mortality, cognitive and physical functioning, and cardiovascular disease (Seeman, McEwen, Rowe, Singer, 2001; Seeman, Singer, Rowe, Horwitz, McEwen, 1997) and various dimensions of social and work environments (Schnorpfeil, Noll, Schulze, Ehlert, Frey, Fischer, 2003; Seeman, Singer, Ryff, Dienberg Love, Levy-Storms, 2002).

Summary measures of physiological dysregulation vary in their structure, underlying assumptions, and the biological components they include. A useful distinction is whether or not they incorporate information on subsequent (tertiary) health outcomes. There are currently two measures in the existing literature that incorporate subsequent health information. The first uses canonical correlation to produce a weighted linear combination of the component biological markers (measured on a continuous scale) that is optimized to predict downstream health outcomes. This measure is positively associated with cognitive and physical decline

(Karlamañgla, Singer, McEwen, Rowe, Seeman, 2002). The second formulation uses recursive partitioning to classify persons into outcome risk categories by iteratively identifying the biological marker (and accompanying cutpoint) that best differentiates among outcomes across individuals. The procedure, which results in pathways characterized by Boolean statements that define high, intermediate and low categories of allostatic load, has provided evidence supporting a link between allostatic load and mortality (Hale, 2004). Singer, Ryff, and Seeman (Singer, Ryff, Seeman, Forthcoming) provide a comparison and discussion of canonical correlation and recursive partitioning, together with an example of a count-based measure.

The use of subsequent health outcomes in the formulation of the summary measure raises two concerns. The first is that it increases the likelihood of overfitting the model underlying the measure; i.e., it may produce a measure that is too specific to the data used to derive it. Its predictive performance in alternative contexts may be impaired. The second issue is that of endogeneity. Summary measures that use information on outcomes are problematic when the measure is derived from, and applied to, the same dataset. For these reasons, we focus on the second group consisting of measures whose formulations do not depend on subsequent (tertiary) health outcomes. We review these measures in the following section.

We know of no systematic examination of measures of allostatic load that 1) investigates variation in the key assumptions governing the calculation of count-based summary measures, and 2) compares the performance of several formulations in explaining and predicting health outcomes. For example, while some studies have used count-based measures that use different biological markers from the original 10 items included in the earliest studies (e.g., Crimmins, Johnston, Hayward, Seeman, 2003; Schnorpfeil, Noll, Schulze, Ehlert, Frey, Fischer, 2003),

these measures are not evaluated relative to the earlier set; the relative contribution of the additional biological variables is unknown.

We use data from a population-based sample of older Taiwanese to compare both the explanatory power and the cross-sectional predictive performance of a range of measures of allostatic load for five health outcomes. We compare several count-based formulations as well as two measures that are fundamentally distinct. None of the measures we examine uses the health outcomes in the calculation of the measure. The count-based formulations comprise indices using 10- and 16- item counts, one- and two-tailed risk categories, and two different percentile cutpoints defining the risk categories. The two distinct measures include one that is based on the standardized distributions of each of the biological measures and one derived from a grade of measurement model that was developed in Seplaki et al. (Seplaki, Goldman, Weinstein, Lin, 2004b).

2 Materials and Methods

2.1 Data

Our data come from the 2000 Social Environment and Biomarkers of Aging Study (SEBAS). SEBAS is based on a longitudinal, nationally-representative probability sample of older Taiwanese—the Taiwan Survey of Health and Living Status. It began in 1989 with persons aged 60 and over and included follow-ups in 1993, 1996, and 1999. A new sample of middle-aged persons (ages 50 to 66) was added for the 1996 wave. As part of SEBAS, a random subsample of 1,713 persons, drawn from both the older and middle-aged cohorts, was selected to be surveyed in 2000 (persons over age 70 in 2000 and those in urban areas were oversampled). This sampling design forms the basis for the two age groups used in this analysis (≤ 70 and >70).

SEBAS consisted of two parts: an in-home survey (N=1,497, a 92% response rate among survivors) and a hospital medical exam conducted by a physician (N=1,023, 68% of those interviewed). Participants in the in-home survey provided data on their health and health history; the subset of medical exam participants also provided fasting blood and 12-hour overnight urine specimens, blood pressure, and anthropometric measurements. Compliance by those completing the medical exam was high (only ten individuals failed to follow the urine protocol, provide a sufficient volume of blood suitable for analysis, or complete the medical exam). Analyses comparing nonparticipants and participants in the medical exam suggest that, in the presence of controls for age, estimates derived from the clinical information are unlikely to be seriously biased (Goldman, Lin, Weinstein, Lin, 2003).

Among the 1,023 persons participating in the medical exam, ten who were missing data on at least one of the biomarkers are excluded from this analysis. An additional fifty-five individuals are excluded because they were missing observations on at least one of the remaining variables, or were identified as having an outlying value for one of the biological measurements (as discussed in greater detail below). These modifications result in a sample of 958 individuals for our analyses.

2.2 Measures

2.2.1 Variables

A total of 16 biological measures, representing both primary mediators and secondary outcomes, are used to construct the various indices examined in this analysis (see Table 2). Primary mediators comprise epinephrine, norepinephrine, dopamine, cortisol, DHEA-S, IGF-1, and IL-6. Secondary outcomes are average systolic and diastolic blood pressure, total serum cholesterol, the ratio of total to high-density lipoprotein (HDL) cholesterol, HDL cholesterol (in

the case of the 10-item index) or triglycerides and total cholesterol (in the case of the 16-item index), fasting glucose, glycosylated hemoglobin (a measure of the percentage of hemoglobin molecules in the blood that are bound to glucose), body mass index (BMI), and waist-to-hip ratio.

Measures are derived from the urine and blood specimens, as well as the physical examination. Twelve-hour urine specimens (necessary to obtain integrated measures of these markers because of their diurnal variation) yielded measures on cortisol, norepinephrine, epinephrine, and dopamine. Measurements for cortisol, epinephrine, norepinephrine, and dopamine are reported as "micrograms per gram creatinine" in order to adjust for body size. The fasting blood specimens yielded assays of HDL cholesterol, total cholesterol, triglycerides, fasting glucose, glycosylated hemoglobin, DHEA-S, IGF-1, and IL-6. The blood pressure and anthropometric measurements collected during the physical examination yielded systolic and diastolic blood pressures (each calculated as the average of two seated blood pressure readings, taken about one minute apart, using a mercury sphygmomanometer with the respondent in a seated position). Data used to calculate BMI (defined as weight in kilograms divided by height in meters squared) and the waist/hip ratio are derived from measures taken for height, weight, and hip and waist circumference.

The health outcomes in this analysis are selected to represent a spectrum of physical and mental functioning reflective of the multi-system physiological dysregulation theoretically associated with allostatic load. Outcomes comprise measures of self-assessed health, impairments in activities of daily living (ADL) and mobility, cognitive performance, and depressive symptoms.

Self-assessed health is measured on a five-point scale (1=poor, 2= not-so-good, 3=average, 4=good, 5=excellent). ADLs comprise bathing, dressing, eating, getting out of bed/standing up/sitting in chair, moving around the house, and going to the toilet. Because of the very low prevalence of ADL impairments in this sample (Table 1), throughout the analysis these six functions are represented by an indicator for the presence of any ADL impairment. Mobility limitations are represented by a count of difficulty squatting, climbing 2-3 flights of stairs, lifting 11-12 kilograms, doing physical work at home, walking 200-300 meters, standing continuously for 15 minutes, and running a short distance (20-30 meters). Depressive symptoms are measured by a 10-item version of the original 20-item Center for Epidemiologic Studies Depression (CES-D) scale (Radloff, 1977). Previous studies have demonstrated that a shortened form of the CES-D yields similar internal consistency, factor structure, and accuracy in detecting depressive symptoms as the full 20-item CES-D among elderly Chinese (Boey, 1999) as well as other populations (Kohout, Berkman, Evans, Cornoni-Huntley, 1993; Shrout, Yager, 1989). Items forming the CES-D index used here (potential range of 0 to 30) include reports in the past week of no interest in eating, sleeping poorly, being in a terrible mood, feeling lonely, people not being nice, feeling anguished, having no energy to do things, feeling joyful (reverse coded), that doing anything is exhausting, and that life is going well (reverse coded). Cognitive function is assessed using a four-item test of temporal orientation (day, month, year, and day-of-the-week) from the modified Short Portable Mental Status Questionnaire (Pfeiffer, 1975). Although the survey includes additional items pertaining to cognitive functioning, these four items are chosen on the basis of evidence suggesting that the subset may be similarly predictive of cognitive impairment as expanded collections, yet less sensitive to confounding by education level (Tractenberg, Weinstein, Weiner, Aisen, Fuh, Goldman, Chuang, Forthcoming).

2.2.2 Calculation of Summary Measures

Calculation of the count-based and GOM score-based indices is accomplished by first dividing each biomarker measurement into three categories: low, middle, and high. The low and high categories are designed to capture values outside optimal operating ranges and the potential for risk at both extremes of each biomarker distribution. To maintain maximum comparability with prior work (e.g., Seeman, Singer, Rowe, Horwitz, McEwen, 1997; Seplaki, Goldman, Weinstein, Lin, 2004b) cutpoints are not defined separately by sex for count-based measures but are sex-specific for the GOM-based measure (see Table 2). Depending on the measure, up to two distinct sets of cutpoints are used to define extreme categories for each of these two sets of measures. These categories are: 1) values below the 10th percentile, between the 10th and 90th percentiles, and above the 90th percentile; and 2) values below the 25th percentile, between the 25th and 75th percentiles, and above the 75th percentile. Several of the lower cutpoints for epinephrine and IL-6 are below assay sensitivity (B.A.S.) (Table 2); assay sensitivity for epinephrine is < 2 µg/L, while assay sensitivity for IL-6 is < 0.1 pg/mL. The one-tailed, count-based index uses only the cutpoints based on quartiles, while alternative versions of the two-tailed and GOM-based measures are each calculated using both sets of cutpoints.

The one-tailed, ten-item count-based index is calculated in the same manner as in the previous literature—i.e., as the number of the following ten biological measurements for which an individual is in the highest risk quartile: cortisol, epinephrine, norepinephrine, DHEA-S, HDL cholesterol, the ratio of total-to-HDL cholesterol, glycosylated hemoglobin, systolic blood pressure, diastolic blood pressure, and the waist-hip ratio (Seeman, Singer, Rowe, Horwitz, McEwen, 1997). The 16-item version of this index is calculated in a corresponding manner but includes an additional five biological measurements—dopamine, IGF-1, IL-6, fasting glucose,

and BMI—and substitutes triglyceride and total cholesterol for HDL cholesterol. In both the ten- and 16-item versions of the one-tailed, count-based formulations, the top quartile (i.e., values above the 75th percentile) of most of the biological measurements is defined as high risk; two exceptions suggested by the medical literature are DHEA-S and HDL cholesterol, where the *bottom* quartile (i.e., values below the 25th percentile) is defined as high risk.

The ten- and 16-item two-tailed, count-based indices are constructed in a similar manner except that for most biological measures values at both extremes are counted as high risk. Exceptions are biological measurements for which the medical literature suggests that only one tail represents elevated risk: DHEA-S (low tail), the ratio of total-to-HDL cholesterol (high tail), and HDL cholesterol (low tail). By redefining risk to include two tails for most parameters, this measure increases the number of individuals classified as high risk for any given biomarker (e.g., given thresholds defined at the 25th and 75th percentiles, half of the sample will be defined as “at risk”). To reduce the proportion of individuals defined as exceeding the risk threshold for the two-tailed count measure, we also use thresholds defined at the 10th and 90th percentiles. We therefore compare four versions of the two-tailed index—corresponding to each combination of the two sets of threshold definitions (10th/90th and 25th/75th percentiles) and the two sets of biological measures (ten and 16 items).

The grade of membership measure is derived and discussed in detail elsewhere (Seplaki, Goldman, Weinstein, Lin, 2004b). Briefly, grade of membership (GOM) (Berkman, Singer, Manton, 1989; Manton, Woodbury, Tolley, 1994; Seplaki, Goldman, Weinstein, Lin, 2004a) is a data-reduction technique. GOM is used here to reduce the large number of variables indicating whether each individual displays low, moderate, or high values on each of 16 biological parameters into a set of individual-specific weights that measure a person's similarity to each of

five, pre-defined, archetypal profiles. Four of the profiles are defined to reflect alternative manifestations of elevated health risk (i.e., low versus high values of primary mediators and low versus high values of secondary outcomes), whereas a fifth “reference” profile is defined to represent the lowest risk (i.e., moderate values on most biomarkers). Thus each individual is assigned a vector of five GOM scores that measures the similarity of the individual’s biomarker indicator values to each respective pure-type profile. Four of the five GOM scores (i.e., excluding the score for the low risk profile) are summed to create a single GOM-based index measuring dissimilarity to the low risk profile. The GOM scores are estimated separately for both sets of biomarker cutpoints (the 10th/90th and 25th/75th percentiles). All GOM analyses reported here are performed using the software by Charpentier (Charpentier, 1996).

The z-score measure is included to provide a formulation that is based on a continuous, rather than a categorical, function of the biological variables. The z-score measure is calculated as the sum of the standardized distances of each of the 16 biological measures from its respective mean. With the exception of DHEA-S and the ratio of total-to-HDL cholesterol, the score is based on both tails (i.e., a score of 1 indicates a value that is 1 SD above OR below the mean). For DHEA-S, a positive z-score is assigned only if the respondent’s value is below the mean; a value of zero is assigned to those values of DHEA-S above the mean because such values are presumed to indicate no health risk. Similarly, the z-score corresponding to the ratio of total-to-HDL cholesterol is positive only for values that are above the mean. After evaluating summary statistics and inspecting the distribution for each biomarker, we identified several extreme values (e.g., all of these outlying values were more than five standard deviations from the mean and most were more than ten standard deviations away) for each of the following biomarkers: dopamine, cortisol, Il-6, ratio total/HDL cholesterol, triglycerides, and fasting glucose. These

outliers (17 for all biomarkers combined) were excluded from the analysis sample before calculating the z-scores.

2.3 Analysis

We perform two sets of analyses that assess the relative explanatory power and cross-sectional predictive performance of each summary measure across the five physical and mental health outcomes. First, the explanatory power of each measure is assessed through comparison of pseudo- R^2 and R^2 statistics. These statistics are derived from separate regression models that predict each of the five health outcomes (each model also includes binary indicators for age > 70 and sex). An ordered probit model is used to predict self-rated health and a logit model is used for the probability of an ADL impairment. Ordinary least squares is used to predict the count of mobility limitations, CES-D, and temporal orientation score.

Second, we evaluate the predictive performance of each measure both within sample and out-of-sample in the following manner. The four non-binary outcomes are first converted to binary indicators: the ordinal measure of self-assessed health is recoded as an indicator of health that is “poor” or “not-so-good” (26%); the count of mobility impairments is converted to an indicator for the presences of any mobility impairment (53%); the CES-D score is recalculated to indicate a score greater than seven (27%) (the conventional cutoff used for the full index is 16 out of a total of 60, so on a proportionate basis eight or more would be the corresponding cutoff for a ten-item CES-D score); and the temporal orientation score is converted to an indicator for a “not normal” score less than four (16%) (Tractenberg, Weinstein, Weiner, Aisen, Fuh, Goldman, Chuang, Forthcoming).

As the next step in evaluating predictive performance, each summary measure of physiological dysregulation is used to predict the five binary health outcomes in separate logit

regression models estimated on the full sample (with binary indicators for age > 70 and sex). Predicted probabilities are generated from each model and used together with the actual (observed) binary outcomes to calculate for each summary measure the area under the receiver operating characteristic (ROC) curve. This approach considers the predicted probability produced from the logit specification as the continuous result from a diagnostic test, for which the actual binary outcome is known. The ROC curve is a graph of the true-positive rate (sensitivity) of this test against its false-positive rate (1-specificity) (Dawson-Saunders, Trapp, 1990). The area under this curve represents the probability that a randomly drawn individual with the outcome will be assigned a higher probability by the diagnostic test (the logit model) than a randomly drawn individual without the outcome (Hanley, McNeil, 1982). Thus a predictive model with an area under the ROC curve equal to 0.5 has no predictive value, while values above 0.5 approaching 1 have increasing levels of predictive value.

Lastly, the out-of-sample predictive performance of each summary measure is assessed by re-estimating the same logit specifications on each of 100 bootstrapped samples of 500 individuals (Efron, Tibshirani, 1993). For each bootstrap sample the predicted probabilities and area under the ROC curve statistics are calculated across the balance of individuals *not* represented in the given sample (thus producing an “out-of-sample” prediction). The ROC curve statistic estimates are then averaged across all 100 bootstrap samples (separately for each health outcome). All analyses (except calculation of the GOM scores, as noted above) were carried out using Stata 8.2 (StataCorp, 2003).

3 Results

The descriptive statistics presented in Table 1 show that 44% of the sample (N=958) is over age 70 and 42% is female. This atypical, male-biased sex ratio reflects the approximately

one million Nationalist military and civilian supporters who migrated to Taiwan from the Mainland in 1949 (Gates, 1981; Tsai, 1992). The mean value for self-assessed health in this sample corresponds to the rating of “average”; very few participants report ADL limitations or score poorly on the temporal orientation battery. Descriptive information for each of the nine measures we evaluate is also listed in Table 1. The percentile cutpoints used to define at-risk values of biological measurements are shown in Table 2. This table reveals substantial variation in percentile values between males and females for some of the biomarkers.

Table 3 displays the R^2 , pseudo- R^2 , and area under the ROC curve statistics (both within and out-of-sample) for each model. Although the estimated model coefficients are not shown, Table 3 shows that only the summed GOM score (based on the 10th and 90th percentiles), the two-tailed 16-item count-based measure (based on the 10th and 90th percentiles), and the z-score index are significantly ($p < 0.05$) associated with all five health outcomes in their respective models (first, seventh, and ninth columns of Table 3, respectively). Conversely, the one-tailed 16-item count-based measure (using quartiles) and the conventional measure used in much of the existing literature (the one-tailed, 10-item measure based on quartiles) are significantly ($p < 0.05$) associated with only the count of mobility limitations (fourth and third columns of Table 3, respectively). These patterns are reflected in the magnitudes of the R^2 and pseudo- R^2 statistics across the summary measures.

In general we do not see very large differences in the predictive performance of the various summary measures, but systematic differences are evident across several of the outcomes—particularly relating to the bootstrapped out-of-sample predictions. For self-assessed health, ADL limitations, and CES-D outcomes, both the z-score and the two-tailed, 16-item measure (using the 10th and 90th percentile cutpoints) typically demonstrate the strongest

performance. Relatively less variability is observed among goodness-of-fit statistics for summary measures of models predicting the counts of mobility limitations and temporal orientation. These results suggest that count-based measures that incorporate both high and low tails of risk appear to be more strongly associated with self-assessed health, ADL limitations, and CES-D than the corresponding measures that incorporate only one tail. This result is stronger for measures that are based on the top and bottom decile cutpoints than the quartile cutpoints.

4 Discussion

A systematic and comprehensive assessment of existing measurement approaches for allostatic load is a critical part of the foundation for developing improved measures. The measures chosen for evaluation in this analysis reflect a wide array of assumptions that have not been tested in the literature. This comparative analysis provides a quantification of these assumptions in terms of relative explained variation and predictive performance. We find that there are modest, but nontrivial effects of the various assumptions underlying measures of allostatic load. For the predictive tasks examined here our findings support the choice of two-tailed, count-based measures relative to one-tailed measures. More restrictive cutpoints for the two-tailed measures (e.g., the 10th and 90th percentiles) perform better than the quartile risk cutpoints generally used in the 10-item count, although it is important to note that decile cutpoints may be less sensitive as an “early warning” threshold for subsequent declines in health and functioning than the quartile cutpoints. More generally, the relatively strong performance of the z-score measure suggests the need for development of measures that preserve the continuous properties of the component biological measurements, while the preference for two-tails over one-tail suggests the importance of nonlinear effects of the biological measurements (Seplaki, Goldman, Weinstein, Lin, 2004a).

There are two limitations of this analysis that should be noted. First, the data used for this evaluation are cross-sectional, so inferences cannot be made regarding the direction of the association between the physiological dysregulation indices and functional outcomes. Second, and perhaps the most significant unresolved issue, is the choice of which biological measurements should be included in summary measures of physiological dysregulation. Two aspects of this issue are the effects on a given summary index of 1) including biological measurements that are not related to physiological dysregulation, and 2) failing to include biological measurements that are related. Inclusion of irrelevant biological measurements would increase the random noise of a summary index. Failure to include important biological measurements would weaken the utility of a summary measure and might occur for two reasons. First, we may not know the “correct” set of measurements to include. Second, we may be unable to measure biological parameters in a manner that is non-invasive and relatively inexpensive.

To the best of our knowledge the present analysis represents the first systematic evaluation of alternative measures for the assessment of multi-system physiological dysregulation associated with allostatic load. We provide several fundamental insights that are of use both to applied researchers in the field who are currently in search of empirical formulations of such constructs, as well as to those who are focused on the development of new measures that improve upon current formulations.

Acknowledgements

This research has been supported by the Demography and Epidemiology Unit of the Behavioral and Social Research Program of the National Institute of Aging, grant numbers R01AG16790 and R01AG16661, and by the National Institute of Child Health and Human Development, grant number 5P30HD32030. We thank Germán Rodríguez for helpful comments.

References

- Berkman, L.F., Singer, B.H., Manton, K., 1989. Black/White Differences in Health status and Mortality Among the Elderly. *Demography*. 26, 661-678.
- Berr, C., Lafont, S., Debuire, B., Dartigues, J.F., Baulieu, E.E., 1996. Relationships of dehydroepiandrosterone sulfate in the elderly with functional, psychological, and mental status, and short-term mortality: a French community-based study. *Proc Natl Acad Sci U S A*. 93, 13410-13415.
- Boey, K.W., 1999. Cross-validation of a short form of the CES-D in Chinese elderly. *Int J Geriatr Psychiatry*. 14, 608-617.
- Charpentier, P., 1996. GOM3, Version 3.4 [Online]. Available by Freeware <http://lib.stat.cmu.edu/DOS/general/.index.html> (verified December 2002).
- Crimmins, E.M., Johnston, M., Hayward, M., Seeman, T., 2003. Age differences in allostatic load: an index of physiological dysregulation. *Exp Gerontol*. 38, 731-734.
- Dawson-Saunders, B., Trapp, R.G., 1990. *Basic and Clinical Biostatistics* Appleton & Lange, Norwalk, Conn.
- Efron, B., Tibshirani, R., 1993. *An introduction to the bootstrap* Chapman & Hall, New York.
- Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. 2001. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *JAMA*. 285, 2486-2497.
- Gates, H., 1981. Ethnicity and Social Class. In: Ahern, E.M., Gates, H., (Eds.), *The Anthropology of Taiwanese Society*. Stanford University Press. pp. 241-281.
- Goldman, N., Lin, I.-F., Weinstein, M., Lin, Y.-H., 2003. Evaluating the Quality of Self-Reports of Hypertension and Diabetes. *J Clin Epidemiol*. 56, 148-154.
- Hale, L., 2004. Life Experiences, Strength of Emotional Response, and Sex-specific Mortality Risk Zones. Office of Population Research Working Paper Series # 2004-01. Princeton University., Princeton.
- Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 143, 29-36.
- Karlamangla, A.S., Singer, B.H., McEwen, B.S., Rowe, J.W., Seeman, T.E., 2002. Allostatic Load as a Predictor of Functional Decline. *MacArthur Studies of Successful Aging. J Clin Epidemiol*. 55, 696-710.
- Kohout, F.J., Berkman, L.F., Evans, D.A., Cornoni-Huntley, J., 1993. Two shorter forms of the CES-D (Center for Epidemiological Studies Depression) depression symptoms index. *J Aging Health*. 5, 179-193.

- Manton, K.G., Woodbury, M.A., Tolley, H.D., 1994. *Statistical Applications Using Fuzzy Sets* John Wiley, New York.
- Mazat, L., Lafont, S., Berr, C., Debuire, B., Tessier, J.F., Dartigues, J.F., Baulieu, E.E., 2001. Prospective measurements of dehydroepiandrosterone sulfate in a cohort of elderly subjects: relationship to gender, subjective health, smoking habits, and 10-year mortality. *Proc Natl Acad Sci U S A.* 98, 8145-8150.
- McEwen, B.S., 2002. Sex, stress and the hippocampus: allostasis, allostatic load and the aging process. *Neurobiol Aging.* 23, 921-939.
- McEwen, B.S., Stellar, E., 1993. Stress and the individual. Mechanisms leading to disease. *Arch Intern Med.* 153, 2093-2101.
- McEwen, B.S., Seeman, T., 1999. Protective and damaging effects of mediators of stress. Elaborating and testing the concepts of allostasis and allostatic load. *Ann N Y Acad Sci.* 896, 30-47.
- Pfeiffer, E., 1975. A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *J Am Geriatr Soc.* 23, 433-441.
- Radloff, L.S., 1977. The CES-D Scale: A Self-report Depression Scale for Research in the General Population. *Applied Psychological Measurement.* 1, 149-166.
- Schnorpfeil, P., Noll, A., Schulze, R., Ehlert, U., Frey, K., Fischer, J.E., 2003. Allostatic load and work conditions. *Soc Sci Med.* 57, 647-656.
- Seeman, T., McEwen, B.S., Rowe, J.W., Singer, B.H., 2001. Allostatic Load As a Marker of Cumulative Biological Risk: MacArthur Studies of Successful Aging. *Proc Natl Acad Sci U S A.* 98, 4770-4775.
- Seeman, T.E., Singer, B.H., Rowe, J.W., Horwitz, R.I., McEwen, B.S., 1997. Price of adaptation--allostatic load and its health consequences. *MacArthur studies of successful aging.* *Arch Intern Med.* 157, 2259-2268.
- Seeman, T.E., Singer, B.H., Ryff, C.D., Dienberg Love, G., Levy-Storms, L., 2002. Social relationships, gender, and allostatic load across two age cohorts. *Psychosom Med.* 64, 395-406.
- Seplaki, C.L., Goldman, N., Weinstein, M., Lin, Y.-H., 2004a. How Are Biomarkers Related to Physical and Mental Well-Being? *J Gerontol A Biol Sci Med Sci.* 59, B201-B217.
- Seplaki, C.L., Goldman, N., Weinstein, M., Lin, Y.-H., 2004b. Measurement of Cumulative Physiological Dysregulation in an Older Population. Office of Population Research Working Paper Series # 2004-02. Princeton University, Princeton, NJ.
- Shrout, P.E., Yager, T.J., 1989. Reliability and Validity of Screening Scales: Effect of Reducing Scale Length. *J Clin Epidemiol.* 42, 69-78.

- Singer, B.H., Ryff, C.D., Seeman, T., Forthcoming. Operationalizing Allostatic Load. In: Schulkin, J., (Ed.), *Allostasis, Homeostasis and the Costs of Physiological Adaptation*. Cambridge University Press, New York, NY. pp. Chapter 4.
- StataCorp. 2003. *Stata Statistical Software: Release 8.2*. Stata Corporation, College Station, TX.
- Svec, F., Lopez, A., 1989. Antiglucocorticoid actions of dehydroepiandrosterone and low concentrations in Alzheimer's disease. *Lancet*. 2, 1335-1336.
- Tractenberg, R.E., Weinstein, M., Weiner, M.F., Aisen, P.S., Fuh, J.-L., Goldman, N., Chuang, Y.-L., Forthcoming. Benchmarking a test of temporal orientation with data from American and Taiwanese persons with Alzheimer's Disease and American normal elderly. *Neuroepidemiology*.
- Tsai, S.-L., 1992. Social Change and Status Attainment in Taiwan: Comparisons of Ethnic Groups. *International Perspectives on Education and Society*. 2, 225-256.

Table 1: Analysis Sample Descriptive Statistics (N=958)

Variable	Mean	Standard Deviation	Minimum	Maximum
<i>Demographic Variables</i>				
Age > 70	0.44	0.50	0	1.00
Female	0.42	0.49	0	1.00
<i>Health Outcome Variables</i>				
Self-Assessed Health	3.08	0.99	1	5.00
Any ADL limitation	0.04		0	1.00
CES-D Score	5.44	5.38	0	28.00
Count of Mobility Limitations	1.66	2.17	0	7.00
Test of Temporal Orientation	3.66	0.88	0	4.00
<i>Physiological Dysregulation Indices</i>				
Summed GOM score (10/90)	0.45	0.24	0	1.00
Summed GOM score (25/75)	0.89	0.17	0.143	1.00
10 item index (75)	2.56	1.59	0	7.00
16 item index (75)	4.04	2.18	0	11.00
Two-tailed 10 item index (25/75)	4.35	1.66	0	9.00
Two-tailed 16 item index (25/75)	7.71	2.19	0	14.00
Two-tailed 10 item index (10/90)	1.90	1.35	0	6.00
Two-tailed 16 item index (10/90)	3.43	1.84	0	10.00
Z-score index (excluding outliers)	10.61	3.31	3.40	26.47

Table 2: Percentile Cutpoints Defining At-Risk Values of Biological Measurements

<i>Biomarker Category</i>		<i>Percentile (N=1023^a)</i>							
		<i>10th</i>	<i>25th</i>	<i>75th</i>	<i>90th</i>				
Cutpoints (sexes combined, used for count-based measures)									
Primary Mediators	Epinephrine (µg/g creatinine) ^b	B.A.S.	0.78	3.67	5.63				
	Norepinephrine (µg/g creatinine)	11.17	15.02	27.09	34.74				
	Dopamine (µg/g creatinine)	87.44	112.35	183.07	226.72				
	Cortisol (µg/g creatinine)	8.75	12.53	29.98	48.01				
	DHEA-S (µg/dL)	20.90	40.80	107.90	152.40				
	IGF-1 (ng/mL)	53.10	69.60	131.70	168.00				
	IL-6 (pg/mL) ^b	B.A.S.	B.A.S.	1.43	3.40				
Secondary Outcomes	Systolic Blood Pressure (mmHG)	114.00	123.00	150.00	166.00				
	Diastolic Blood Pressue (mmHG)	70.00	75.00	90.00	97.00				
	Total Cholesterol (mg/dL)	153.00	175.00	225.00	252.00				
	Ratio of Total Cholesterol to HDL	2.81	3.40	5.11	6.14				
	Triglycerides (mg/dL)	54.00	71.00	147.00	204.00				
	HDL Cholesterol (mg/dL)	33.00	39.00	57.00	67.00				
	Fasting Glucose (mg/dL)	84.00	89.00	107.00	138.00				
	Glycosylated Hemoglobin	4.80	5.10	5.80	7.10				
	BMI	19.98	22.00	26.56	28.87				
Waist/Hip Ratio	0.80	0.84	0.93	0.96					
Cutpoints by Sex (used for GOM-based measure)		<i>Males (N=590^a)</i>				<i>Females (N=433^a)</i>			
		<i>10th</i>	<i>25th</i>	<i>75th</i>	<i>90th</i>	<i>10th</i>	<i>25th</i>	<i>75th</i>	<i>90th</i>
Primary Mediators	Epinephrine (µg/g creatinine) ^b	B.A.S.	0.85	3.41	4.90	B.A.S.	0.63	4.09	6.65
	Norepinephrine (µg/g creatinine)	10.42	13.89	24.23	32.91	12.88	17.71	30.30	36.16
	Dopamine (µg/L)	80.24	102.29	158.02	196.07	104.11	133.63	207.82	254.73
	Cortisol (µg/g creatinine)	7.87	11.47	25.40	43.04	10.06	14.51	34.35	53.61
	DHEA-S (µg/dL)	32.00	53.45	125.30	172.10	13.00	29.20	78.20	118.00
	IGF-1 (ng/mL)	54.30	73.45	137.15	173.70	49.80	66.80	121.30	151.90
	IL-6 (pg/mL) ^b	B.A.S.	B.A.S.	1.40	3.10	B.A.S.	B.A.S.	1.50	3.80
Secondary Outcomes	Systolic Blood Pressure (mmHG)	113.00	123.00	149.00	164.00	114.00	125.00	152.00	170.00
	Diastolic Blood Pressue (mmHG)	69.50	75.00	90.00	96.00	70.00	74.00	90.00	97.00
	Total Cholesterol (mg/dL)	148.00	170.00	218.00	242.00	163.00	182.00	232.00	263.00
	Ratio of Total Cholesterol to HDL	2.77	3.40	5.21	6.19	2.86	3.35	4.98	5.98
	Triglycerides (mg/dL)	52.00	67.00	139.00	199.00	60.00	79.00	155.00	209.00
	Fasting Glucose (mg/dL)	84.00	88.00	105.00	127.00	85.00	89.00	113.00	164.00
	Glycosylated Hemoglobin	4.80	5.10	5.70	6.50	4.90	5.10	6.10	8.10
	BMI	19.66	21.78	26.03	28.20	20.44	22.37	27.11	30.38
	Waist/Hip Ratio	0.84	0.87	0.94	0.97	0.78	0.81	0.90	0.94

^aCounts represent the maximum number of observations over which the percentiles for any one of the biomarkers are calculated; the number of missing observations varies across the biomarkers.

^bThreshold values for epinephrine and IL-6 are below assay sensitivity (B.A.S.); approximately 33% of readings on IL-6 and 20% of readings on epinephrine fell below the sensitivity of the assays (< 2 µg/L for epinephrine and < 0.1 pg/mL for IL-6).

Table 3: Comparison of Model Fit and Prediction Statistics (N=958)^a

Health Outcome	Prediction Error Measure	Physiological Dysregulation Score (percentile cutpoint)								
		Summed GOM score (10/90)	Summed GOM score (25/75)	10 item index (75)	16 item index (75)	Two-tailed 10 item index (10/90)	Two-tailed 10 item index (25/75)	Two-tailed 16 item index (10/90)	Two-tailed 16 item index (25/75)	Z-score index (excluding outliers)
<i>Self-Assessed Health</i>										
	Pseudo-R ²	0.012**	0.011**	0.009	0.009	0.009	0.009	0.011**	0.010*	0.013**
	Area Under ROC Curve	0.591	0.591	0.581	0.583	0.586	0.581	0.592	0.583	0.603
	Out-of-Sample Area Under ROC Curve	0.567	0.569	0.554	0.561	0.564	0.557	0.575	0.563	0.585
<i>ADL Limitations</i>										
	Pseudo-R ²	0.075**	0.053	0.041	0.048	0.069**	0.039	0.091**	0.050	0.082**
	Area Under ROC Curve	0.716	0.693	0.661	0.678	0.707	0.667	0.740	0.692	0.745
	Out-of-Sample Area Under ROC Curve	0.663	0.628	0.610	0.624	0.652	0.597	0.695	0.633	0.693
<i>CES-D Score</i>										
	R ²	0.044**	0.032**	0.026	0.026	0.040**	0.030*	0.044**	0.034**	0.046**
	Area Under ROC Curve	0.630	0.627	0.606	0.608	0.620	0.618	0.637	0.628	0.637
	Out-of-Sample Area Under ROC Curve	0.618	0.610	0.584	0.590	0.603	0.599	0.620	0.610	0.621
<i>Mobility Limitations</i>										
	R ²	0.205**	0.192**	0.199**	0.197**	0.196**	0.193**	0.206**	0.194**	0.218**
	Area Under ROC Curve	0.746	0.744	0.752	0.747	0.736	0.747	0.739	0.746	0.750
	Out-of-Sample Area Under ROC Curve	0.738	0.735	0.747	0.740	0.727	0.741	0.731	0.737	0.744
<i>Test of Temporal Orientation</i>										
	R ²	0.083*	0.081	0.082	0.082	0.082	0.079	0.084*	0.083*	0.084*
	Area Under ROC Curve	0.721	0.720	0.719	0.720	0.713	0.716	0.720	0.724	0.723
	Out-of-Sample Area Under ROC Curve	0.712	0.704	0.701	0.710	0.700	0.705	0.711	0.711	0.710

^aAll models include binary indicators for age (>70) and sex, in addition to the measure of interest. Pseudo-R² statistics are derived from the ordered probit (self-assessed health) and logit (ADL limitations) specifications, while the R² statistics are derived from the OLS (CES-D score, mobility limitations, and test of temporal orientation) specifications. The area under ROC curve statistics are derived from logit models predicting binary transformations of each health outcome—as noted in the text, these are: self-assessed health “not-so-good” or “poor”; any ADL limitations, CES-D score > 7; any mobility limitations; test of temporal orientation < 4. The out-of-sample area under ROC curve statistics are estimated as the average area under the ROC values over 100 bootstrapped samples of 500 individuals; predicted values for each sample are calculated only for individuals not in the bootstrap sample.

* Estimated regression coefficient (not shown, available upon request) is significant at 5%, or ** significant at 1%. All standard errors are estimated using the Huber-white robust variance estimator.