A Comparison of Biomarkers Across Two Older Populations: Assessment of Consistency and Generalizability

Christopher L. Seplaki[1]
Noreen Goldman[2]
Maxine Weinstein[3]
Tara Gruenewald[4]
Arun S. Karlamangla[4]
Teresa E. Seeman[4]

DRAFT—Preliminary and incomplete. Please do not cite or distribute.

Draft version 9/21/06 submitted for PAA 2007

[1]Center on Aging and Health and Department of Population and Family Health Sciences, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD
[2]Center for Health and Wellbeing and Office of Population Research, Princeton University, Princeton, NJ
[3]Center for Population and Health, Georgetown University, Washington, D.C.
[4]Division of Geriatrics, Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA

Abstract

Researchers continue to expand the array of biomarkers of physiological system function collected alongside traditional self-reported measures in population-based social surveys. Still, little is known about the variability of many biomarkers across populations or the consistency of associations between biomarkers and self-reported health measures. We take steps to address this gap using a wide array of traditional clinical and non-traditional biomarkers of physiological system function, together with two self-reported health outcomes (one physical, one cognitive), from two large, comparable, population-based studies of older adults from the U.S. and Taiwan. Our results highlight basic features of distribution variability and predictive consistency. We find that, while many biomarkers are similar, there is enough variability to warrant caution on the part of researchers, depending on the type of biomarker and the association under investigation.

Abstract word count = 131

1 Introduction

In 2001 the National Academy of Sciences published the landmark volume *Cells and Surveys* (Finch, Vaupel and Kinsella 2001). The chapters discussed the myriad issues stemming from the increasing combination of self-report and biological assessment techniques in population research. The charge of the report comprised the, "what, why, whether, who, and how" related to collecting biological information in social surveys (Finch and Vaupel 2001, p.2). This influential volume continues to serve as an important resource for population researchers whose investigations merge self-reported survey information with detailed physical exams and biological specimens (e.g., blood, saliva, urine). Researchers can use such physical information to produce biomarkers[1] of a diverse array of physiological, anthropometric, or genetic characteristics or processes than can be used to predict downstream health outcomes (i.e., clinical endpoints), to understand individual disease etiology, and to identify mechanisms that drive health-related social phenomena or changes in population health (e.g. see the discussion by (Crimmins and Seeman 2001).

Since 2001 many large, population-based social surveys in the U.S. and around the world have analyzed numerous biomarkers from different biological sources, and researchers continue to expand the array of biomarkers that are collected from large samples. The most common biomarkers are anthropometric measures (e.g., height, body mass index (BMI), waist-hip ratio). Other biomarkers are traditional indicators of disease risk that are commonly used in clinical practice (e.g., cardiovascular disease risk indicators such as blood pressure or serum cholesterol). Still other biomarkers include laboratory measurements that are neither typical diagnostic

[1] The term biomarker is often assumed to mean *biomarker of aging*; here we speak of biomarkers more generally as indicators of any biological process or physiological system. McClearn (1992, 1997) discusses concerns regarding basic properties of biomarkers of aging.

indicators nor measured routinely in large numbers of individuals (e.g., neuroendocrine markers of stress, inflammation, and genotypes of disease risk).

However, researchers face several obstacles when adopting biomarkers into population-based research. Chief among these is our limited understanding of the population distribution of many biomarkers, particularly for those that are relatively new or adapted from small, laboratory samples or settings. Users of such measures face a number of fundamental analytic issues for which there is often little guidance from the literature, such as the reasonableness of assumptions regarding distributions, the application of statistical transformations and categorizations, and the identification of risk groups. Laboratory procedures can also be a source of important variation in biomarker data. There is very little research examining such variation in biomarkers across populations or on the consistency of associations between biomarkers and outcomes across populations. Uncertainty surrounding these basic characteristics means that we can say little about the generalizability of both the measures and their associations with outcomes. This uncertainty imposes serious limitations on many population studies that use biomarkers and affects the confidence we have in the broad application of their findings.

We address this critical gap in the literature by comparing a wide array of both traditional clinical and non-traditional hormonal biomarkers of physiological system functioning from two large, population-based studies of older adults from the U.S. and Taiwan. We test our hypotheses (below) through analysis of the distributions of biomarkers across samples and evaluations of the consistency of each biomarker in predicting two distinct health outcomes, one physical and one cognitive, both within sample and out-of-sample. We assess the implications of the results for generalizability and provide guidance to researchers in the field for use of these measures in future studies.

Few studies examine variation in biomarkers across populations; those that do primarily examine markers of cardiovascular disease risk. A study by Cai et al. (2004) compares the levels of total cholesterol, systolic blood pressure, and BMI across population-based samples of adults in their 30's to 60's from four countries (China, US, Poland, Russia). They find that both BMI and total cholesterol were lowest in the sample from China, for both men and women. Systolic blood pressure was lower in China than the US for men, but roughly equal for women. Banks et al. (2006) show higher glycosylated hemoglobin levels in U.S. older adults compared with English counterparts, higher systolic blood pressure and lower levels of HDL cholesterol. Though not comparing across populations, Crimmins et al. (2005) study changes in ten biomarkers between over-65 cohorts in the U.S. approximately eight years apart through the 1990s. These include systolic and diastolic blood pressures, HDL, total and fasting LDL cholesterol, glycosylated hemoglobin, and BMI. Cholesterol measures improved, while blood pressure and BMI worsened. Goldman et al. (2004) use data for 10 biomarkers drawn from the same two studies we examine here (and also add a second U.S. sample) to compare non-traditional, hormonal biomarkers (epinephrine, norepinephrine, cortisol, and DHEA-S) as well as traditional measures of cardiovascular disease risk (HDL cholesterol, ratio of total to HDL cholesterol, systolic and diastolic blood pressures, glycosylated hemoglobin, and waist-hip ratio). Focusing primarily on contrasting differences in biomarkers between men and women within-sample, they find that patterns are distinct for each set of biomarkers. They identify smaller differences in cardiovascular health measures between sexes in Taiwan than in the U.S., fueled in part by more at-risk values for U.S. men (relative to U.S. women), and in part by more at-risk values among Taiwanese women (relative to Taiwanese men). In contrast, women in both countries generally have worse values of the hormonal biomarkers relative to the men in their

respective countries. Goldman et al. (2004) do also look across samples, and point out higher waist-hip ratio values among U.S. men relative to Taiwanese men (and lower values for U.S. versus Taiwanese women), higher values for glycosylated hemoglobin, epinephrine, and norepinephrine in the U.S. (within both sexes), and somewhat higher blood pressure readings among Taiwanese versus U.S. women (p.399).

We identified only one study that examines variability in associations between biomarkers and outcomes across populations. The study by Cai et al. (2004) demonstrates variability in the relationship between total cholesterol and total mortality. The relationship of cardiovascular-related mortality with total cholesterol was consistent across three of the countries, but only for men (China was not included in this part of their analysis).

Within a given study population, a few researchers (using the same samples we analyze here) have examined associations between various downstream outcomes and biomarkers, focusing on the contributions of the non-traditional, hormonal biomarkers relative to the traditional measures of cardiovascular disease risk. Karlamangla et al. (2002) study the same U.S. data that we examine and demonstrate a significant contribution of hormonal biomarkers in predicting functional decline independent from that of biomarkers of cardiovascular disease risk. In an analysis of mortality in the Taiwanese data (below), Turra et al. (2005) show that biomarkers from both traditional and non-traditional biomarkers are significant predictors.

We evaluate the following three hypotheses. First, we hypothesize that the sex-specific distributions of each biomarker should be similar across the two samples, with differences in our controlled comparison consistent with those noted by Goldman et al. (2004). Second, we hypothesize that the associations observed between the outcomes and the biomarkers will generally be consistent across both samples. Lastly, taking the predicted probability generated

by a given model as reflecting the net effect of relationships within that model, we hypothesize that out-of-sample predictions (prediction of outcomes using data from one sample but model parameters from the other) will be comparable to the relevant within sample predictions.

2 Methods

2.1 Data

Data come from two population-based studies of older adults in the U.S. (the MacArthur Study of Successful Aging, MSSA) and Taiwan (the Social Environment and Biomarkers of Aging Study, SEBAS). Each study includes a baseline collection of biomarkers derived from blood and 12-hour urine specimens and a follow-up comprising self-reported measures of physical function and interviewer-administered cognitive function items. As described below, the studies differ in several important ways, including sampling frame, baseline year, and the duration between baseline and follow-up. The steps taken to address these differences and maximize the comparability of the samples are discussed in the following sections.

The MSSA began in 1988 with a community-based sample of men and women aged 70 to 79 drawn from three cohorts of the NIA's Established Populations for Epidemiologic Studies of the Elderly (EPESE) (Durham NC, East Boston MA, and New Haven CT)(Berkman et al. 1993; Cornoni-Huntlev et al. 1986). MSSA participants were screened based on their age and four physical and two cognitive function measures to represent roughly the top third of their age group. Of the 4,030 individuals in EPESE, 1,313 satisfied the screening criteria and 1,189 (90.6%) of these agreed to participate in the MSSA and are the baseline cohort.

The baseline assessment for the MSSA comprise a face-to-face interview and a physical examination. The examination included blood pressure readings, anthropometric measurements,

a blood draw, and a 12-hour overnight urine sample. Although baseline examinations are available for over 900 subjects, only 729 have complete blood and urine data. Analyses comparing these 729 subjects with the complete cohort suggest that they are representative of the full cohort (Seeman et al. 1997). Follow-up interviews completed between October 1995 and February 1997 provide measures of self-reported health and physical function. Of the 1,189 individuals alive at the completion of the baseline interview, 202 had died by follow-up, 722 completed face-to-face interviews, 107 had proxy-partial interviews, and 87 either refused or were alive but not successfully contacted.

The sampling frame for the 2000 SEBAS is based on the longitudinal study of older Taiwanese, the Taiwan Survey of Health and Living Status (see Goldman, Glei and Chang (2003) for details). This prior study identified a nationally representative probability sample of persons aged 60+ in 1989. Baseline interviews for the study were conducted in 1989 and reinterviews followed in 1993, 1996, 1999, and 2003. A new cohort of persons aged 50 to 66 was incorporated for the 1996 wave, and both the new and the original cohort were interviewed for the 1999 and 2003 waves. The 2000 SEBAS comprised a random subsample of 1,713 persons from both the older and middle-aged cohorts in 1999 (the SEBAS design also oversampled persons aged 70+ in 2000 and those in urban areas). The 2000 SEBAS included both 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).

Using a protocol similar to the MSSA, the subset of medical exam participants in SEBAS provided blood pressure readings, anthropometric measurements, a blood draw, and a 12-hour overnight urine sample. Findings from Goldman et al. (2003) suggest that bias may be limited in age-adjusted comparisons of the clinical information in SEBAS between the medical exam

participants and nonparticipants. Measures of self-reported health and function measures come from the 2003 follow-up interview of the Taiwan Survey of Health and Living Status. Of the original 1,023 SEBAS hospital medical exam participants, 72 died by 2003, 937 were alive as of 2003, and the survival status of 14 participants is unknown.

Subsequent modifications to each sample (described below) and exclusions of observations with missing information, yield total analysis samples for both outcomes of 268 for SEBAS and 359 for MSSA. Analysis samples specific to each outcome are 139 and 235 for the SEBAS mobility and cognitive impairment logit models, respectively, and 259 and 215 for the MSSA mobility and cognitive impairment models, respectively.

2.2 Measures

The MSSA and SEBAS studies each provide data on 14 biomarkers that reflect a broad range of physiological function. The collection of this specific set of biomarkers is motivated by the conceptual model of *allostatic load,* which describes the development of dysregulation across multiple physiologic systems (HPA-axis, sympathetic nervous, immune, and cardiovascular) that is driven by the cumulative effects of stressful experiences across the life course (McEwen 1998; McEwen and Wingfield 2003).

Over half of the biomarkers are common clinical diagnostic indicators that signal risk for a variety of deleterious outcomes (e.g., cardiovascular disease). These include systolic and diastolic blood pressures, high-density lipoprotein (HDL) cholesterol, the ratio of total to HDL cholesterol, glycosylated hemoglobin, body mass index (BMI), waist-to-hip ratio, and creatinine clearance. The blood pressure measures in each survey are averages of two seated readings using a mercury sphygmometer. Glycosylated hemoglobin is a measure of blood sugar (specifically, it is the percentage of hemoglobin molecules in the blood that are bound to glucose

and thus reflects sugar control over a longer time period than does fasting glucose). In addition to providing information on renal function, creatinine clearance is often used (as in the present study) to normalize hormone values derived from urine specimens for body size.

The remaining 6 biomarkers are not generally collected in clinical settings and do not have widely recognized clinical cutoff points for risk. These nontraditional biomarkers are hormones that measure the activity of several physiological systems, including the HPA-axis, and sympathetic nervous, and immune systems. Cortisol, epinephrine and norepinephrine are considered “stress hormones” because they are integral to the range of physiologic responses to stressful stimuli (Brown 1994). Dopamine serves multiple functions in the brain and, in particular, may play a part in coordinating responses to stress (Nieoullon and Coquerel 2003). Dehydroepiandroster-one sulfate (DHEA-S) is a steroid that has received increasing attention in the medical literature as a predictor of positive outcomes (Glei and Goldman 2006). Interleukin-6 (IL-6) is a hormonal marker of inflammation; increasing levels of IL-6 are associated with adverse outcomes (Leng et al. 2004).

In both the SEBAS and MSSA the nocturnal resting measures of cortisol, norepinephrine, epinephrine, dopamine, and creatinine clearance come from the 12-hour, overnight urine specimens. We adjust the measures of cortisol, norepinephrine, epinephrine, and dopamine for body size by dividing by the amount of creatinine in the urine (i.e., each is presented in units of "micrograms per gram creatinine"). Creatinine clearance is calculated from plasma and urine creatinine concentrations and urine rate (volume divided by the period of collection). Assays of HDL cholesterol, total cholesterol, glycosylated hemoglobin, DHEA-S, and IL-6 come from the blood specimens; the blood pressure and anthropometric measurements come from the physical examination. BMI is equal to weight in kilograms divided by height in meters squared.

The health outcome information recorded at follow-up in each study includes reports of mobility limitations and a cognitive performance evaluation. Table 1 describes each health outcome measure in detail and compares them across the two surveys. Cognitive function is assessed using items adapted from the modified Short Portable Mental Status Questionnaire (SPMSQ)(Pfeiffer 1975) that are consistent between the two surveys. The mobility limitation indicators are the same for each study, though the number of response categories differs (four for the SEBAS and between two and five for the MSSA). The steps taken to maximize comparability of variables across samples in each of biomarkers and outcomes are described below.

3 Analysis

Following the perspective similar of Goldman et al. (2004), we evaluate our hypotheses primarily through focus on the inferences derived from each sample and not on formal statistical tests of differences between samples. However, before performing the analyses, we take several steps to maximize the comparability of the SEBAS and MSSA samples. We first restrict the SEBAS sample to individuals aged 70-80 and drop from the analysis individuals in the SEBAS with at least one limitation in activities of daily living at baseline, reflecting the sampling screener applied in the MSSA study. We also transform the original biomarker values to rankits, replacing each value by the corresponding quantile of a normal distribution. Specifically, we recode the i^{th} ordered observation of each variable with the expected value of a standard normal random variable truncated to lie between the value immediately above and below the value of interest, using the formulae in Johnson, Kotz and Balakrishnan (1994) (p.156). This transformation serves two important purposes. First, it removes any bias that would be introduced by potential variation in laboratory measurements over time and between studies

(there is an approximately ten year difference in when the laboratory assays were completed for the MSSA and SEBAS). Second, the rankit transformation removes the influence of outliers in the biological data. The rankit transformation does remove issues presented by censored data (e.g., observations for biomarkers such as IL-6 or epinephrine that fall below assay sensitivity). Lastly, because our interest is in the incidence of mobility limitations and cognitive difficulty, analysis is restricted to those who haven no difficulty or impairment at baseline. All analyses are performed using Stata Version 9.1 (StataCorp 2005).

Conceptually, our analyses can be divided into two parts. The first part consists of a descriptive comparison of individual biomarkers across the SEBAS and MSSA samples and comprises several diagnostic evaluations. We assess the distributional characteristics of each of the biomarkers by calculating basic summary statistics, perform the chi-squared test for normality by D'Agostino, Belanger and D'Agostino Jr. (1990, combining evaluations of skewness and kurtosis)[2] and then examine two sets of plots. The first set displays the raw biomarker values against their rankit transformation, i.e., normal probability plots. Because the rankit transformations are (by definition) normally distributed, biomarkers that are normally distributed will yield approximately linear plots; non-normal distributions will not and the shape of the distributions reveals insights into their characteristics (e.g., skewness), as illustrated by D'Agostino et al. (1990, Figure 3, p.321). We draw the normal probability plots separately for each study. The second set of diagnostic plots display the quantiles of a given biomarker for one sample against the other, i.e., a quantile-quantile plot, facilitating direct comparison of the distributions of each biomarker across studies. Because the focus in these plots is on the actual values of each biomarker (not rankits) and shows both samples on the same graph, we generate plots separately for men and women.

[2] We implement the normality test using Stata's *sktest* command with *noadust* option.

The second part of the analysis studies the practical impact of observed differences in biomarkers across the two samples by 1) examining consistency of the associations of the biomarkers with each of the two outcomes, and 2) analyzing the within and out-of-sample predictive performance of the biomarkers. We recalculate each health outcome as a binary indicator (i.e., an indicator of any mobility limitation and an indicator for at least one incorrect cognitive evaluation response) in order to simplify the prediction exercise and enable use of a logit specification. We first estimate separate logit models for each biomarker to predict each outcome (a noted previously, the analysis is restricted to those who did not have either limitation at baseline). Logit models include only the rankit transformation of one biomarker, the square of the rankit (to capture nonlinear associations), and controls for age at baseline and sex. In order to address confounding across biomarkers, we next estimate a combined biomarker model for each sample that simultaneously includes all 14 biomarkers and their associated nonlinear terms, together with controls for age and sex. Because of the large number of terms in these specifications and the limited number of observations, we bootstrap each of the combined models. This yields mean parameter estimates and bias corrected standard errors calculated over 5000 repetitions. We do not apply weights to the SEBAS to correct for the sample design because our interest is on a specific subset of the SEBAS sample and applying our results to the MSSA (we are not concerned with extrapolating the results of these models to the SEBAS population). Lastly, we contrast within and out-of-sample predicted probabilities from each of the full models. Each model is estimated again using 5000 repetitions and predicted probabilities of the binary outcomes are calculated using both data and model parameters from one sample, and then using data from one sample with model parameters from the other. For each bootstrapped analysis, bias corrected 95% confidence intervals are calculated[3].

[3] This calculation performed using Stata's *bootstrap* command.

4 Results

We provide descriptive information for the raw values of each variable in Table 2. Since the specific number of observations in each sample differs across the two health outcomes, summary statistics for age, sex and the biomarkers are shown for the combined number of observations (SEBAS N=268 and MSSA N=359). Summary statistics for the binary health outcomes are given for the relevant subsamples. Table 2 demonstrates that the average age at baseline for each sample is the same (after restricting the SEBAS to the 70-80 range). Inspection of the two binary self-reported health outcomes shows that MSSA participants report higher proportions of each, which likely reflects the longer follow-up period of the MSSA (seven versus three years for SEBAS).

The biomarker summary information in Table 2 shows comparable mean values for many of the biomarkers, but also highlights the broad ranges that exist for several of the measures, particularly the neuroendocrine biomarkers. Almost every biomarker fails the formal, combined test for normality, which is not surprising given the potential for skewness suggested by the broad ranges demonstrated for most (relative to their standard deviations), e.g., dopamine and IL-6. In addition, not observable in the table (but detailed in the footnotes) is the fact that 65 values (24% of 268) for epinephrine and 91 values (34% of 268) for IL-6 in the combined SEBAS sample are below assay sensitivity, and therefore recorded as “zeros”. Also not investigated in Table 2 are potential differences in the biomarkers across samples by sex.

Figures 1a and 1b provide additional insights using normal probability plots, displaying raw biomarker values against their normally-distributed rankit transformations for the SEBAS (Figure 1a) and MSSA (Figure 1b) samples. The first result is that the corresponding plots for each biomarker are very similar across the two studies. In both Figures 1a and 1b, the

nonlinearity of most plots provides evidence against normality for those biomarkers; this result is particularly notable among the non-traditional measures. Specifically, the shape of the plots suggests skewness of the distributions characterized by long (right) tails. The figures also demonstrate that many biomarker distributions also contain outliers, some of which are extreme. For example, the plots for dopamine and IL-6 in Figure 1a (SEBAS) required exclusion of two extreme observations each so that the balance of distribution could be shown. Among the traditional biomarkers, another result common to both studies is the marked nonlinearity (i.e., non-normality) of the glycosylated hemoglobin distributions.

Figures 2a and 2b enable direct comparison of biomarker values across samples by plotting the quantiles of each biomarker in the SEBAS against those in the MSSA; the solid line in each graph represents the 45° diagonal (i.e., across study equality of values). Because the sex distribution differs between the two studies (approximately 31% of the 268 total SEBAS sample is female and 50% of the 369 total MSSA is female), Figure 2a displays results for men while 2b displays results for women. Results for men (Figure 2a) demonstrate that the quantile values across studies for average systolic blood pressure, creatinine clearance, cortisol, and DHEA-S are approximately equivalent (abstracting from the cortisol outlier), shown by values that fall roughly along the diagonal. Values for average diastolic blood pressure, HDL cholesterol are generally higher in SEBAS men, while values for glycosylated hemoglobin, BMI, waist-to-hip ratio, norepinephrine, dopamine, and IL-6 are higher in MSSA men. The difference between samples also increases with the biomarker values for dopamine, IL-6, and norepinephrine. The distribution of values for total-to-HDL cholesterol ratio is slightly higher in the MSSA, particularly among higher values of the ratio (ignoring the lone outlier). The large number of values below assay sensitivity for epinephrine in the SEBAS exaggerates the apparent difference

between SEBAS and MSSA men in this biomarker at the low end of the distribution; values are roughly comparable through the balance of the distribution.

Biomarker distributions for women are less consistent across samples than those for men and female inequalities across studies generally stem from higher values among SEBAS women than MSSA women. Only quantiles for HDL cholesterol, creatinine clearance, and DHEA-S are approximately equivalent across the range of values. Values for average systolic and diastolic blood pressures, waist-to-hip ratio, and cortisol are generally higher in SEBAS women, while values for dopamine, IL-6, and norepinephrine are consistently higher in MSSA women. Like the men, the differences between samples for these last three biomarkers also increase roughly with their values (suggesting either a nonlinear relationship or possible differences in laboratory technology). Comparisons across studies of the distributions for glycosylated hemoglobin, BMI, and epinephrine are more difficult to interpret (and reinforce the added value of this detailed analysis beyond that provided by the summary statistics in Table 2). Like the men, values for glycosylated hemoglobin are higher in the MSSA, but the difference is smaller for women. Unlike the men, female BMI values are approximately equal across samples for values in the lower half of their BMI distribution, while the upper half of the distribution is somewhat higher among MSSA women. Like the men, SEBAS women also reflect many epinephrine values below assay sensitivity, exaggerating difference between SEBAS and MSSA women at the low end of the distribution; values are roughly comparable through the balance of the distribution.

The second part of our analysis investigates the potential effects that differences in biomarkers across samples may have on estimates of relationships between biomarkers and frequently examined outcomes. Figure 3 displays the coefficient estimates and 95% confidence intervals from twenty-eight separate logit models (one for each biomarker and sample), each

predicting the presence of any mobility limitation at follow-up (among those with no limitations at baseline). Independent variables in each model are the rankit transformation of the biomarker, the square of the rankit (to capture nonlinear associations), and controls for age at baseline and sex. Each plot pairs the linear terms from each sample next to each other for contrast, and similarly the squared terms.

Figure 3 shows, with only the rank order of biomarkers preserved, significant associations in both samples of the probability of any mobility limitation at follow-up with BMI (positive association) and with creatinine clearance (negative association). However, Figure 3 also shows several differences between samples in the associations between biomarkers and this outcome. Linear terms are different for dopamine, where we find a strong negative association in the SEBAS sample but not the MSSA. We see differences in marginally significant linear terms for glycosylated hemoglobin (positive association in SEBAS), DHEA-S (negative association in SEBAS) and waist-hip ratio (positive association in MSSA). Inferences from the squared terms differ only for BMI, where the positive association between BMI and the probability of any mobility limitation at follow-up in SEBAS is nonlinear.

Figure 4 shows the results of a similar exercise conducted using a different outcome: at least one incorrect response to the cognitive evaluation questions in Table 1. We find no significant associations that are consistent across both samples for this outcome. Inferences from the linear terms differ again for waist-hip ratio (positive association in MSSA) and, though marginally significant, again for glycosylated hemoglobin (positive association in SEBAS) and DHEA-S (negative association in SEBAS), and (newly) for BMI (positive association in MSSA). Inferences regarding nonlinear relationships differ between the samples for glycosylated

hemoglobin and epinephrine, where the quadratic term is at least marginally significant in the SEBAS but not the MSSA.

To address confounding across biomarkers, we provide a comprehensive analysis in Figures 5 and 6, where we enter all biomarkers (and their quadratic terms) together into a single model, with controls for age at baseline and sex. Figure 5 displays the bootstrapped parameter estimates and 95% confidence intervals (bias corrected) for the mobility outcome, while Figure 6 shows results for the cognitive outcome. Figure 5 shows that, in contrast to the individual models, significant associations are only observed in the MSSA sample. Specifically, in the MSSA we find that mobility limitation is positively associated with BMI and negatively associated with creatinine clearance, as found in the individual models. The point estimates from the SEBAS model are uniformly much less precise than those from the MSSA model, given the available sample sizes. The models in Figure 6 reveal no significance associations for either study population (each specification has similar numbers of observations available). However, alternatively, we also derive no conflicting inferences from across studies on associations between the cognitive outcome and the biomarkers from the models in Figure 6.

Our final set of results in Figure 7 test the practical effects that differences in the parameter estimates found in the combined models actually have on extrapolation of predicted outcome probabilities from each model to the other sample. The goal of this exercise is to examine if a predicted probability (which reflects the aggregate effect of all relationships in the underlying model) calculated within sample (i.e., the parameters are those produced from the same data) can be extrapolated to another, independent sample and yield a comparable result. The left side of the figure shows results for the mobility outcome and the right side shows the cognitive outcome results. For each outcome, four probabilities (and 95% confidence intervals)

are calculated; the two probabilities on the left (A/B and E/F) use the parameter estimates from the combined SEBAS model and the two on the right (C/D and G/H) use the parameter estimates from the combined MSSA model (the same two models that produced the bootstrapped parameter estimates shown in Figure 5). Lastly, within each set of estimates (for each outcome), the predicted probability on the left (A and C, E and G) uses the SEBAS data and the predicted probability on the right (B and D, F and H) uses the MSSA data.

Figure 7 shows that the predicted probabilities are not significantly different for the mobility limitation outcome. Within sample estimates (A vs. D) overlap, though suggest a higher probability of poor outcome in the MSSA (possibly reflecting to some degree the longer follow-up in the MSSA). Out-of-sample predictions are also consistent; applying the SEBAS estimates to the MSSA data (A vs. B) yields comparable probabilities, as does applying the MSSA estimates to the SEBAS data (D vs. C). We draw a similar conclusion for the cognitive outcome, thought the results differ slightly. Within sample estimates (E vs. H) do not overlap, showing a significantly higher mean probability of poor outcome in the MSSA (again, we note the longer follow-up in the MSSA). Out-of-sample predictions are not significantly different from their within sample counterparts, however. Applying the SEBAS estimates to the MSSA data yields a predicted probability similar to the within sample estimate (E vs. F). Similarly, applying the MSSA estimates to the SEBAS data also generates a probability comparable to the within sample estimate (H vs. G). In sum, the within and out-of-sample predicted probabilities, representing the net effect of all model coefficients together, demonstrate that this effect is generalizable out of sample for the specified models and outcomes.

5 Discussion

Our analyses evaluate the characteristics of a wide array of biomarkers from two population-based studies representing both Western and non-Western older adults. We partially confirm our first hypothesis that the sex-specific distributions of each biomarker should be similar across the two samples (and any differences consistent with those noted by the unrestricted comparison by Goldman et al. (2004)). We find that many biomarkers display similar characteristics across samples; virtually all fail a formal parametric test for normality. Although non-normal (e.g., skewed) distributions often suggest a need for special analytic methods, Lumley et al. (2002) point out that this is often not necessary for large samples. However, we find several pronounced differences between the US and Taiwanese participants beyond those noted in the prior work by Goldman et al. (2004), and our analyses provide additional detail about the sources of variation. For example, our findings in Figures 2a and 2b for glycosylated hemoglobin, epinephrine, and norepinephrine show that elevated levels in the U.S. are concentrated at particular points in each distribution. Our analysis in Figures 1a and 1b also highlight the presence of outliers, which may embody true variation or reflect laboratory errors or other technical (nonbiological) sources of variation. A third example is provided by the monotonic increases in sample differences (for both men and women) evident for dopamine, IL-6 and norepinephrine shown in Figures 1a and 1b. While these patterns may reflect true variability across the samples, it is also possible that systematic differences exist in laboratory procedures between the two studies.

Our results for BMI are consistent with those of Cai et al. (2004), demonstrating higher values in the MSSA sample than the SEBAS sample (though we find that values are roughly comparable for women at lower levels of BMI). Though we did not examine total cholesterol

specifically, our results for HDL cholesterol and the total-to-HDL cholesterol ratio are also consistent with their findings (we show generally more healthy levels of each in the SEBAS). Our results for systolic blood pressure are not consistent with their findings; we find approximately equal values for men in our samples, and Taiwanese women show slightly higher values than American women (also shown by Goldman et al. (2004)) . The difference in blood pressure findings could be attributable to the younger study participants examined by Cai et al. (2004), given the complexity of the relationship between blood pressure and age (for a discussion of this relationship see Arking (2006), pp. 64-67).

We also show that knowledge about consistency (or inconsistency) for a given biomarker across samples from one sex does not imply similar generalizability for the other sex. Specifically, Figures 2a and 2b show that comparability across samples differs between men and women for five biomarkers: average systolic blood pressure, HDL cholesterol, waist-hip ratio, cortisol, and to a lesser degree BMI. Unfortunately, this group contains some of the most frequently analyzed biomarkers in the literature. This finding suggests that assertions regarding consistency of a biomarker distribution across different samples for, say, men, should not be extended to that biomarker for women (and vice versa). We also find that, in general, the biomarker distributions for women are less consistent across samples than those for men and, in further contrast with the results for men, female differences across studies generally stem from higher values among SEBAS women rather than the MSSA women (a finding in line with those by Goldman et al. (2004)). In sum, our results suggest that not only might biomarkers differ between populations (within sex), but that these differences may not be consistent across men and women, and that the variation is not limited to nontraditional, hormonal biomarkers but includes traditional, anthropometric measures as well.

Our second hypothesis described our expectations for consistency across studies in associations that we would see with two outcomes, one physical and one cognitive. Both the single association models (Figures 3 and 4) and combined models (Figures 5 and 6) provide mixed support for our initial hypotheses. In support of our hypothesis, we do not find significant associations that are contradictory between the two studies (e.g., significant relationships in opposite directions); in fact, specific point estimates (ignoring significance) are most often consistent across studies. This finding agrees with those for men by Cai et al. (2004). However, Figures 3 and 4 suggest a greater number of significant associations in the Taiwanese sample than the MSSA, and the majority of these pertain to the traditional biomarker measures, not the hormonal indicators. Yet subsequent results from the combined models reveal that the single association model findings for SEBAS are not robust to inclusion of additional confounders. Specifically, Figure 5 shows that only the associations of incident mobility limitation with BMI (positive association) and creatinine clearance (negative association) in MSSA are robust to controls for the other biomarkers. Again, this may reflect a difference in the association between the studies, differences in sample sizes and statistical power between the studies, or possibly differing strength of a common mechanism over distinct time spans.

Our final hypothesis addresses a core issue facing this and related research; namely, despite differences in biomarkers across samples, and variation observed in the associations between biomarkers and outcomes from sample to sample, how might these differences affect prediction of outcomes in other samples or populations? We hypothesized that out-of-sample predictions (prediction of outcomes using data from one sample but model parameters from the other) will be comparable to the within sample prediction. The analysis we present in Figure 7 supports this hypothesis for both outcomes, improving our confidence in the generalizability of

predictive results (that are "net" of individual associations) produced from population-based biological measures.

We note several limitations of our analysis. We first raise a conceptual point. As stated at the outset, our goal is *not* to investigate in detail associations between biomarkers and outcomes (there are extensive literatures devoted to such evaluations), but instead to examine variability of association in a controlled way—using identical methods on data that we transform to be as similar as possible across studies. This implies that emphasis shift away from the significance (or lack of significance) of the majority of the associations we estimate, toward the *consistency or inconsistency* of the findings *across* samples. A related point is that many of the associations we estimate are likely affected by the small sample sizes we analyze; however, again, the sample sizes are comparable and should yield comparable power. To the extent that the SEBAS samples are somewhat smaller, this may affect the relative numbers of significant associations that we report. However, throughout the discussion we emphasize the direction of the associations, rather than their statistical significance, in order to limit the effects of differences in power on our conclusions. Also, other transformation strategies may yield different results for the associations; our preference in choosing to analyze the rankit transformation was to provide the highly conservative results imparted by preserving only the rank order of biomarker values across studies. This approach has the disadvantage of removing variability, and thus power, in samples that are already limited in size. Second, some of the differences in associations with outcomes may be due to the difference in length of follow-up. The association between a biomarker and the incidence of either mobility difficulties or cognitive impairment may indeed differ by, in this case, whether assessment of the outcome

occurs at 3 or 7 years[4]. In addition, the different follow-up times may also affect differential selection on the samples over time; i.e., assuming those in poorer health are more likely to die, such selection has 7 years over which to influence the MSSA and only 3 for the SEBAS. We also note that additional attrition occurred in both samples from refusal, loss to follow-up, and missing information. Like mortality, such sample modification has the potential to alter the health profile of the sample, thus in our analysis we implicitly assume that health profiles of missing or dropped individuals mirror those included in the analysis. A similar issue arises with respect to the observations for epinephrine and IL-6 in the SEBAS that are below assay sensitivity. Finally, our test of out-of sample consistency is very conservative because a mean probability, by definition, will obscure potentially interesting variation across different subgroups, for example, men and women. We intend for our result in that analysis to be illustrative, with a deeper, focused analysis of generalizability properties left for future work. Detailed analysis of the distributions of predicted probabilities along the lines conducted in the first part of our analysis here, and separated by sex, would be particularly enlightening.

Several strengths of our study are also worth highlighting. First, in addition to the use of comparable biomarkers and longitudinal data on outcomes, our sample focuses on older adults—a population for which the study of health is especially challenging and the use of biomarkers holds exceptional promise. This is because older populations typically reflect complex mixes of comorbidity that confound study of single endpoints or pathways. Older populations may also reflect unmeasured, senescent mechanisms that can influence relationships between predictive factors and outcomes (Austad 2001). Second, our sample compares a Western sample with a non-Western sample. Because the majority of research that uses biomarkers comes from

[4] We are in the process of obtaining 3 year follow-up data for the MSSA to revise the analysis and remove this limitation.

Western populations, we know little about how our biomarkers might vary from sample to sample and they may be influenced by cultural differences. We cannot untangle the biological from the cultural in our analysis. But we can suggest that, before findings in the literature using biomarkers can be generalized, researchers need to develop analytic plans that address the potential for systematic, as well as random, variability in biomarker measurement.

The addition of biomarkers to population studies holds great promise to deepen our understanding of the effects of social factors on health, which can inform both population and biological research. However, expansion of the range of measures has tended to outpace our understanding of fundamental, population characteristics of such measures. Our results fill an important gap by highlighting basic features of a broad range of biomarkers that are increasingly featured in population studies. We find that, while many biomarkers are similar, there is enough variability to warrant caution on the part of researchers, depending on the nature of the association being investigated.

Acknowledgements

Preliminary work on this research was supported by the NIA Center for Demography of Aging at the Center for Health and Wellbeing, Princeton University, 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 would like to thank Germán Rodríguez and Burton Singer for statistical advice.

Corresponding author: Christopher L. Seplaki, PhD, Center on Aging and Health, The Johns Hopkins University, 2024 E. Monument Street, Suite 2-700, Baltimore, MD 21205. E-mail: cseplaki@jhsph.edu.

References

Arking, R. 2006. *The biology of aging: observations and principles*. Oxford; New York: Oxford University Press.

Austad, S.N. 2001. "Concepts and Theories of Aging." Pp. 3-22 in *Handbook of the Biology of Aging, 5th Edition*, edited by E.J. Masoro and S.N. Austad. New York: Academic Press.

Banks, J., M. Marmot, Z. Oldfield, and J.P. Smith. 2006. "Disease and disadvantage in the United States and in England." *Jama* 295(17):2037-2045.

Berkman, L.F., T.E. Seeman, M. Albert, D. Blazer, R. Kahn, R. Mohs, C. Finch, E. Schneider, C. Cotman, G. McClearn, and et al. 1993. "High, usual and impaired functioning in community-dwelling older men and women: findings from the MacArthur Foundation Research Network on Successful Aging." *Journal of Clinical Epidemiology* 46(10):1129-1140.

Brown, R.E. 1994. *An Introduction to Neuroendocrinology*. Cambridge; New York: Cambridge University Press.

Cai, J., A. Pajak, Y. Li, D. Shestov, C.E. Davis, S. Rywik, Y. Li, A. Deev, and H.A. Tyroler. 2004. "Total cholesterol and mortality in China, Poland, Russia, and the US." *Ann Epidemiol* 14(6):399-408.

Cornoni-Huntlev, J., D.B. Brock, O. A.M., T. J.O., and W. R.B. 1986. *Established populations for the epidemiologic studies of the elderly: resource data book. NIH Publication No. 86-2443*. Bethesda, MD: NIH.

Crimmins, E.M., D. Alley, S.L. Reynolds, M. Johnston, A. Karlamangla, and T. Seeman. 2005. "Changes in biological markers of health: older Americans in the 1990s." *J Gerontol A Biol Sci Med Sci* 60(11):1409-1413.

Crimmins, E.M.and T.E. Seeman. 2001. "Integrating Biology into Demographic Research on Health and Aging (With a Focus on the MacArthur Study of Successful Aging)." Pp. 9-41 in *Cells and Surveys: Should Biological Measures Be Included in Social Science Research?* edited by C.E. Finch, J.W. Vaupel, and K. Kinsella. Washington, D.C.: National Academy Press.

D'Agostino, R., A. Belanger, and R. D'Agostino Jr. 1990. "A suggestion for using powerful and informative tests for normality." *The American Statistician* 44(4):316-321.

Finch, C.E.and J.W. Vaupel. 2001. "Collecting Biological Indicators in Household Surveys." Pp. 1-8 in *Cells and Surveys: Should Biological Measures Be Included in Social Science Research?* edited by C.E. Finch, J.W. Vaupel, and K. Kinsella. Washington, D.C.: National Academy Press.

Finch, C.E., J.W. Vaupel, and K.G. Kinsella. 2001. "Cells and surveys: should biological measures be included in social science research?" Washington, D.C.: National Academy Press, National Research Council (U.S.). Committee on Population.

Glei, D.A.and N. Goldman. 2006. "Dehydroepiandrosterone Sulfate (DHEAS) and Risk for Mortality Among Older Taiwanese." *Ann Epidemiol*.

Goldman, N., D.A. Glei, and M.-C. Chang. 2003. "The Role of Clinical Risk Factors in Understanding Self-Rated Health." *Annals of Epidemiology* 13(5):1-9.

Goldman, N., I.-F. Lin, M. Weinstein, and Y.-H. Lin. 2003. "Evaluating the Quality of Self-Reports of Hypertension and Diabetes." *Journal of Clinical Epidemiology* 56(2):148-154.

Goldman, N., M. Weinstein, J. Cornman, B. Singer, T. Seeman, and M.C. Chang. 2004. "Sex differentials in biological risk factors for chronic disease: estimates from population-based surveys." *Journal of Women's Health* 13(4):393-403.

Johnson, N.L., S. Kotz, and N. Balakrishnan. 1994. *Continuous univariate distributions*. New York: Wiley & Sons.

Karlamangla, A.S., B.H. Singer, B.S. McEwen, J.W. Rowe, and T.E. Seeman. 2002. "Allostatic Load as a Predictor of Functional Decline. MacArthur Studies of Successful Aging." *Journal of Clinical Epidemiology* 55(7):696-710.

Leng, S.X., A.R. Cappola, R.E. Andersen, M.R. Blackman, K. Koenig, M. Blair, and J.D. Walston. 2004. "Serum levels of insulin-like growth factor-I (IGF-I) and dehydroepiandrosterone sulfate (DHEA-S), and their relationships with serum interleukin-6, in the geriatric syndrome of frailty." *Aging Clin Exp Res* 16(2):153-157.

Lumley, T., P. Diehr, S. Emerson, and L. Chen. 2002. "The importance of the normality assumption in large public health data sets." *Annu Rev Public Health* 23:151-169.

McClearn, G.E. 1992. "The reliability and stability of biomarkers of aging." *Ann N Y Acad Sci* 673:1-8.

—. 1997. "Biomarkers of age and aging." *Exp Gerontol* 32(1-2):87-94.

McEwen, B.S. 1998. "Protective and damaging effects of stress mediators." *New England Journal of Medicine* 338(3):171-179.

McEwen, B.S.and J.C. Wingfield. 2003. "The concept of allostasis in biology and biomedicine." *Horm Behav* 43(1):2-15.

Nieoullon, A.and A. Coquerel. 2003. "Dopamine: a key regulator to adapt action, emotion, motivation and cognition." *Curr Opin Neurol* 16 Suppl 2:S3-9.

Pfeiffer, E. 1975. "A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients." *Journal of the American Geriatric Society* 23(10):433-441.

Seeman, T.E., B.H. Singer, J.W. Rowe, R.I. Horwitz, and B.S. McEwen. 1997. "Price of adaptation--allostatic load and its health consequences. MacArthur studies of successful aging." *Archives of Internal Medicine* 157(19):2259-2268.

StataCorp. 2005. "Stata Statistical Software: Release 9.1." College Station, TX: StataCorp LP.

Turra, C.M., N. Goldman, C.L. Seplaki, D.A. Glei, Y.-H. Lin, and M. Weinstein. 2005. "Determinants of Mortality at Older Ages: The Role of Biological Markers of Chronic Disease." *Population and Development Review* 31(4):677-701.

Table 1: Comparison of self-reported health outcome variables from the SEBAS and MSSA

Variable	**Additional Description: SEBAS**	**Additional Description: MSSA**
Mobility Limitations		
	Walk 200-300 meters, Use stairs, Work around home, Raise hands over head, Squat, Grasp/ turn objects with fingers, Lift 11-12kgs	Walk 1/2 mile, Use stairs, Work around home, Raise hands above shoulder, Stoop/ crouch/ kneel, Write/handle small objects, Lift over 10 lbs.
Cognition (Items from the SPMSQ)		
	Today's date (month, day, year), Day of week, Mother's maiden name, Current President, President before him, Serial 3 calculation	Today's date, Day of week, Mother's maiden name, Current President of the U.S., President before him, Serial 3 calculation

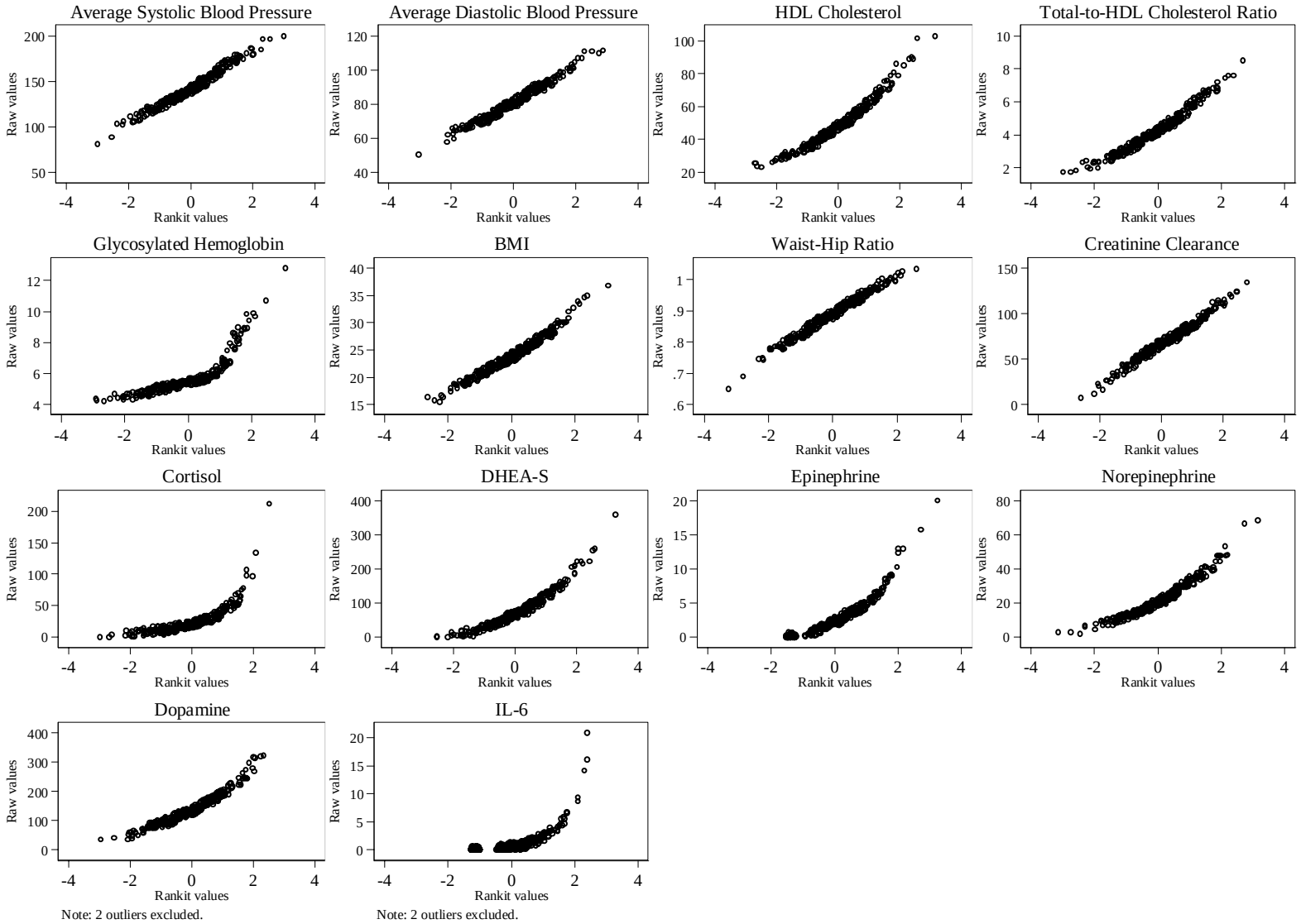
Table 2: Sample summary information[a]

	SEBAS					MSSA				
	Mean	**SD**	**Min**	**Max**	**Test for normality ($Pr>\chi^2$)[d]**	**Mean**	**SD**	**Min**	**Max**	**Test for normality ($Pr>\chi^2$)[d]**
Age at baseline (years)	74.0	2.8	70.0	80.0		74.0	2.7	70.0	80.0	
Female	0.31		0.0	1.0		0.50		0.0	1.0	
Raw biomarker values										
Average Systolic Blood Pressure (mmHG)	142.6	19.8	83.0	200.0	0.34	137.4	18.5	97.0	198.0	0.00
Average Diastolic Blood Pressure (mmHG)	81.4	10.2	50.0	111.0	0.08	76.6	9.8	50.0	109.0	0.02
HDL Cholesterol (mg/dL)	49.6	14.0	25.0	103.0	0.00	47.7	14.4	10.0	108.0	0.00
Total-to-HDL Cholesterol Ratio	4.3	1.2	1.7	8.7	0.00	5.0	1.7	2.1	19.2	0.00
Glycosylated Hemoglobin (% of Hb)	5.7	1.2	4.3	12.8	0.00	6.7	1.8	4.1	20.2	0.00
BMI	23.9	3.5	16.2	37.2	0.00	26.0	4.1	14.5	42.1	0.00
Waist-Hip Ratio	0.9	0.1	0.6	1.0	0.10	0.9	0.1	0.6	1.1	0.00
Creatinine Clearance (mL/min)	70.0	21.3	8.0	136.2	0.44	68.7	24.2	20.5	152.0	0.00
Cortisol (ul/g creatinine)	24.8	21.5	3.1	216.8	0.00	21.9	16.2	0.7	112.6	0.00
DHEA-S (ug/dL)	77.3	51.4	0.0	360.9	0.00	71.0	47.4	5.0	295.0	0.00
Epinephrine (ul/g creatinine)[c]	2.6	2.8	0.0	19.9	0.00	3.8	1.9	0.9	11.3	0.00
Norepinephrine (ul/g creatinine)	21.6	10.5	1.6	70.9	0.00	40.2	22.3	1.7	220.0	0.00
Dopamine (ul/g creatinine)	242.5	1341.2	24.3	20868.7	0.00	253.8	100.5	6.4	858.3	0.00
IL-6 (pg/mL)[c]	2.2	13.0	0.0	164.4	0.00	3.9	4.4	0.6	36.5	0.00
Binary health outcome indicators at followup	**N**	**Mean**	**Min**	**Max**		**N**	**Mean**	**Min**	**Max**	
Any mobility limitation	139	0.51	0.0	1.0		259	0.64	0.0	1.0	
At least one incorrect cognitive evaluation response	235	0.30	0.0	1.0		215	0.55	0.0	1.0	

[a]SEBAS N=268 and MSSA N=359 which is equal to the combined number of observations across both health outcomes in each sample, respectively. Sample sizes for specific binary health outcomes are noted in the bottom two rows of the table.
[b]Followup data are approximately 3 years after baseline for the SEBAS and 7 years after baseline for the MSSA.

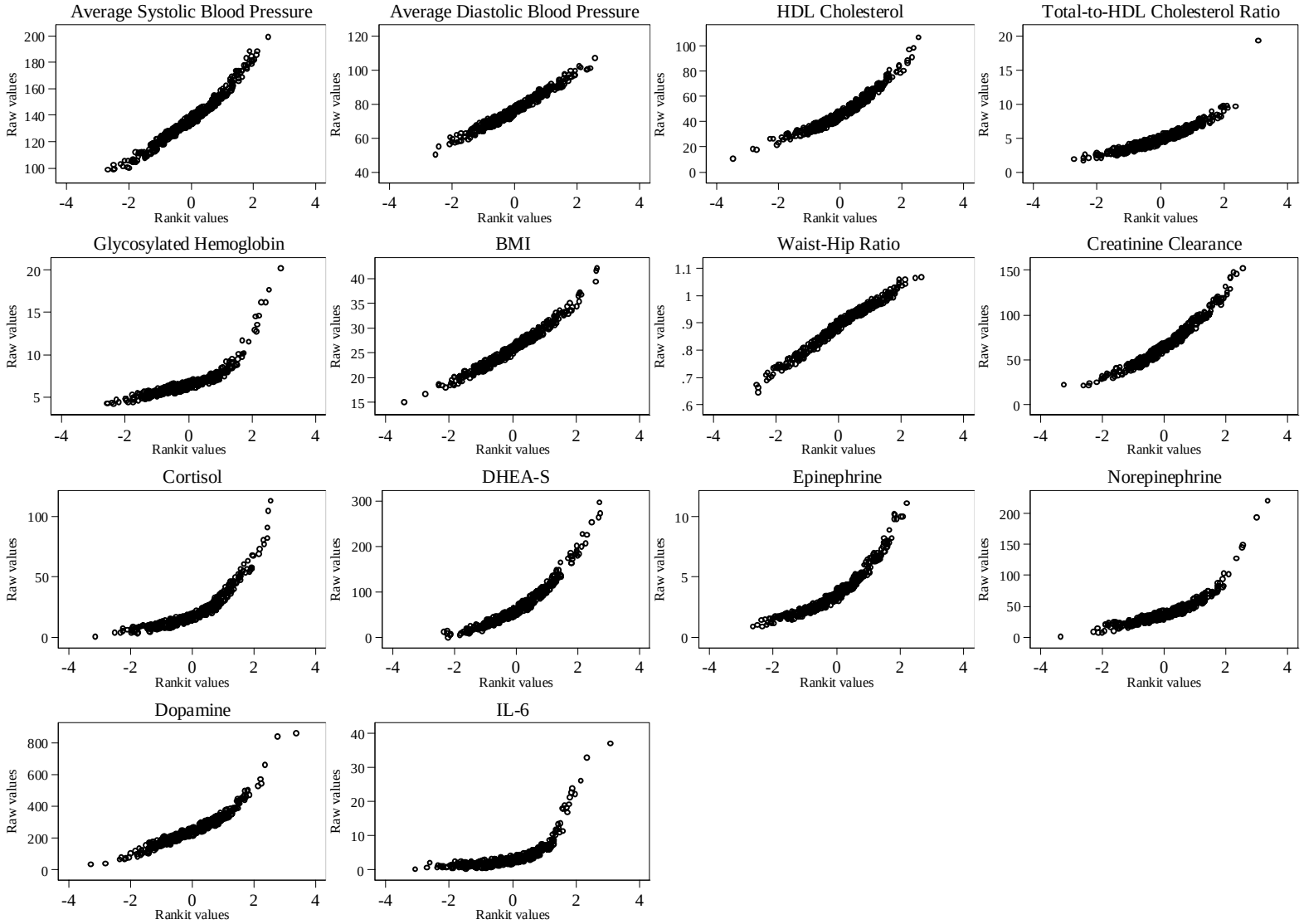
[c]65 values (24%) for epinephrine and 91 values (34%) for IL-6 in the combined SEBAS sample are below the sensitivity of the assays (equal to 2 μg/L for epinephrine and 0.1 pg/mL for IL-6).
[d]Normality test is the combined χ^2 test for skewness and kurtosis by (D'Agostino et al. 1990) and calculated using Stata's *sktest* command with *noadust* option.

Figure 1a: Raw biomarker values versus rankit values, SEBAS[a]



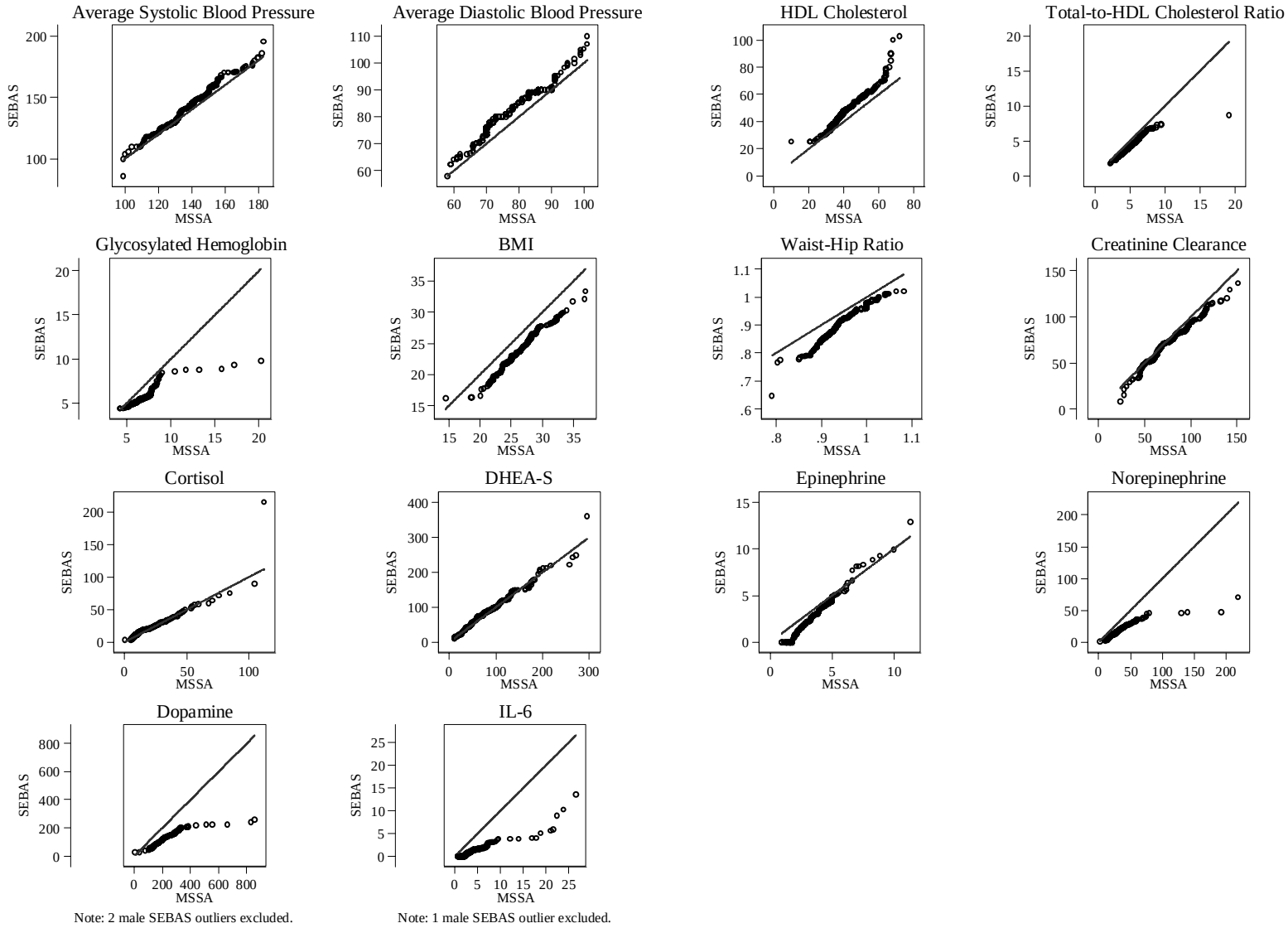
[a]Sample N= 268 for each plot (except dopamine and IL-6, N=266 as noted), which is equal to the combined number of observations across both health outcomes. Random noise (spherical) added to all points to assist in visualization of overlapping points.

Figure 1b: Raw biomarker values versus rankit values, MSSA[a]



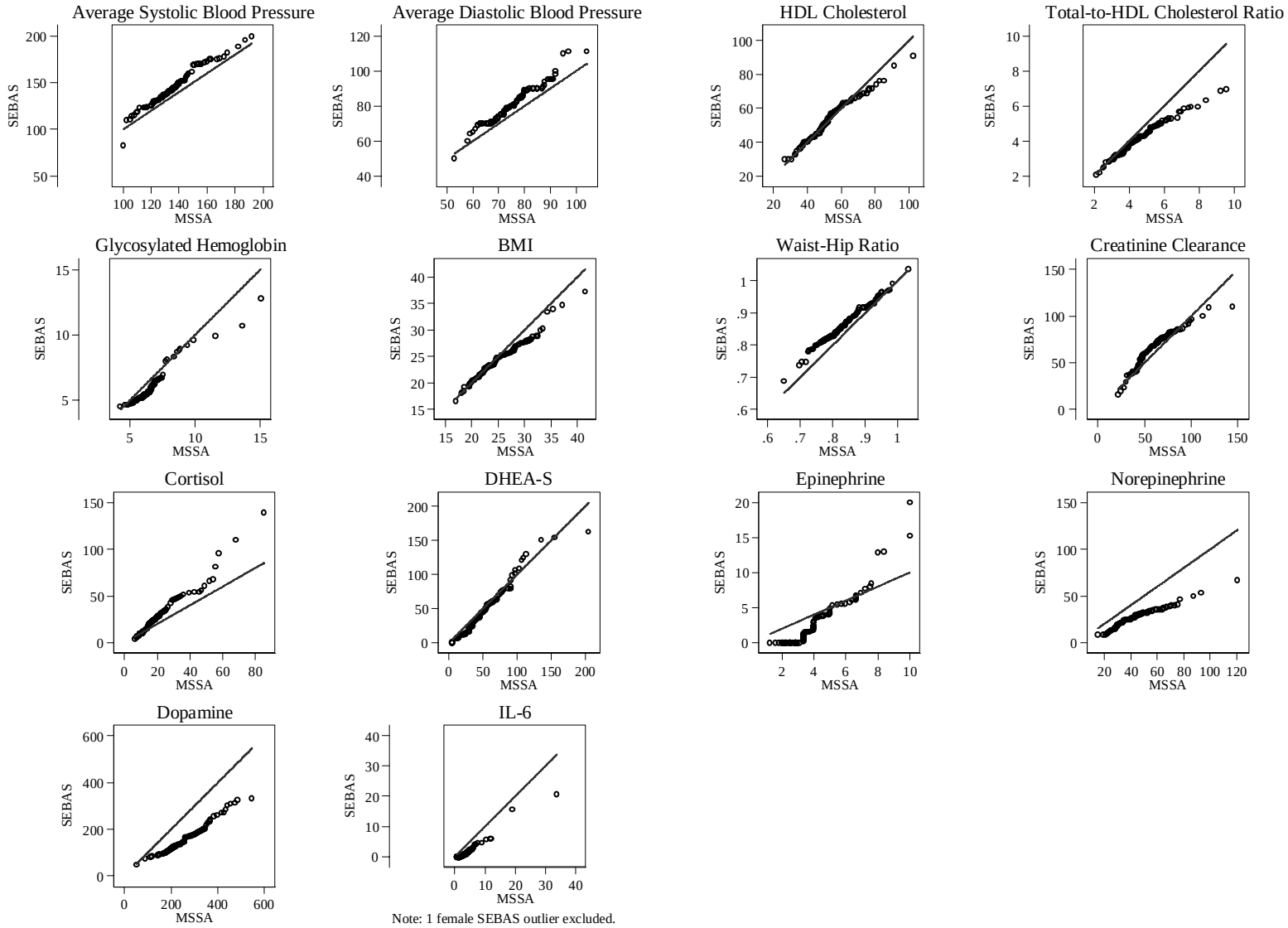
[a]Sample N=359 for each graph, which is equal to the combined number of observations across both health outcomes. Random noise (spherical) added to all points to assist in visualization of overlapping points.

Figure 2a: Quantile-Quantile plots of biomarkers in both the SEBAS and MSSA studies, men[a]



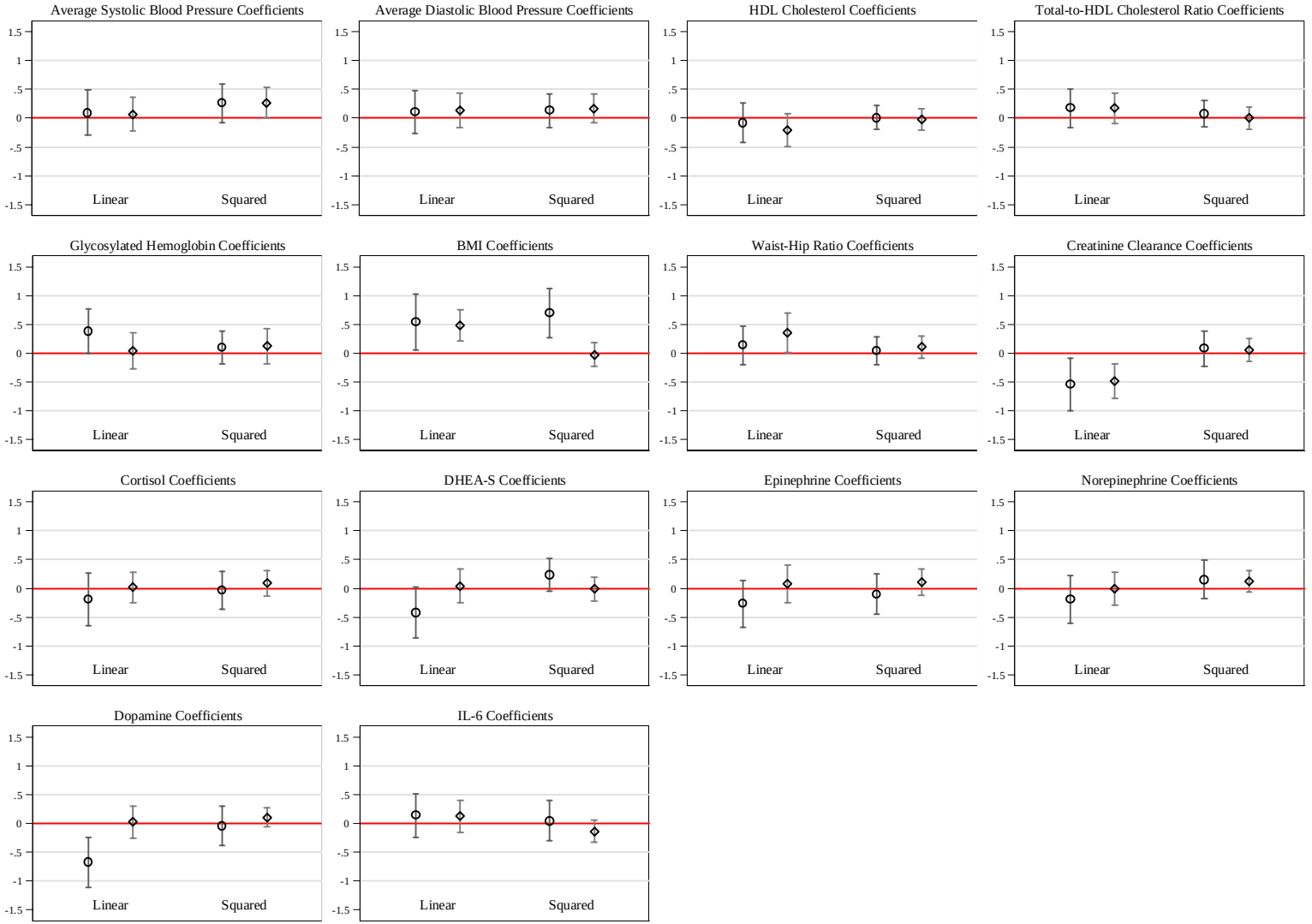
[a]For each graph N = 186 men (out of 268 total) for SEBAS and N = 180 men (out of 359 total) for MSSA, representing the combined number of male observations across both health outcomes in each sample. (N=184 and 185 for dopamine and IL-6 in SEBAS, as noted.)

Figure 2b: Quantile-Quantile plots of biomarkers in both the SEBAS and MSSA studies, women[a]



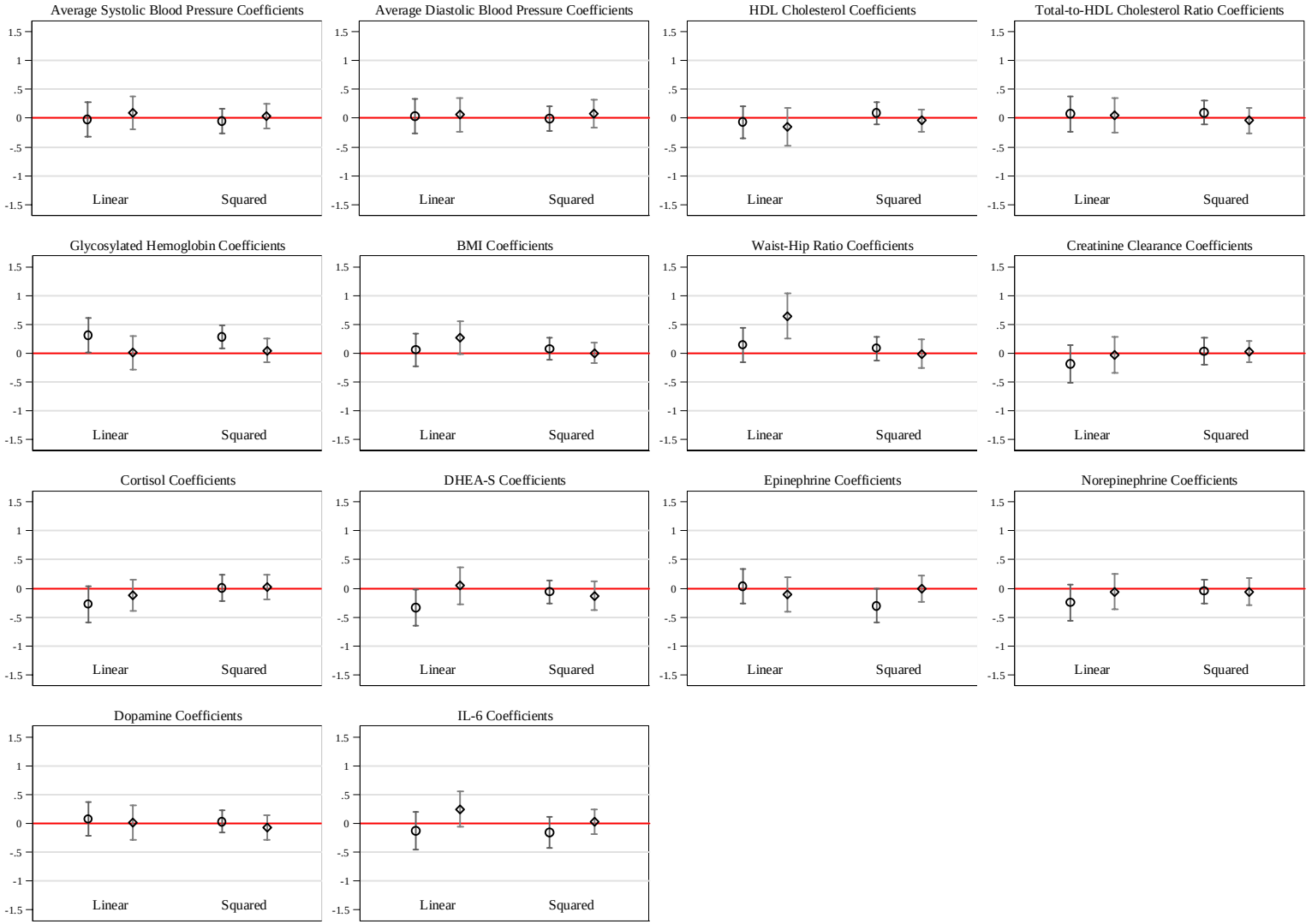
[a]For each graph N = 82 women (out of 268 total) for SEBAS and N = 179 women (out of 359 total) for MSSA, representing the combined number of female observations across both health outcomes in each sample. (N=81 for IL-6 in SEBAS, as noted.)

Figure 3: Parameter estimates from separate logit models predicting any mobility limitation at follow-up[a,b]



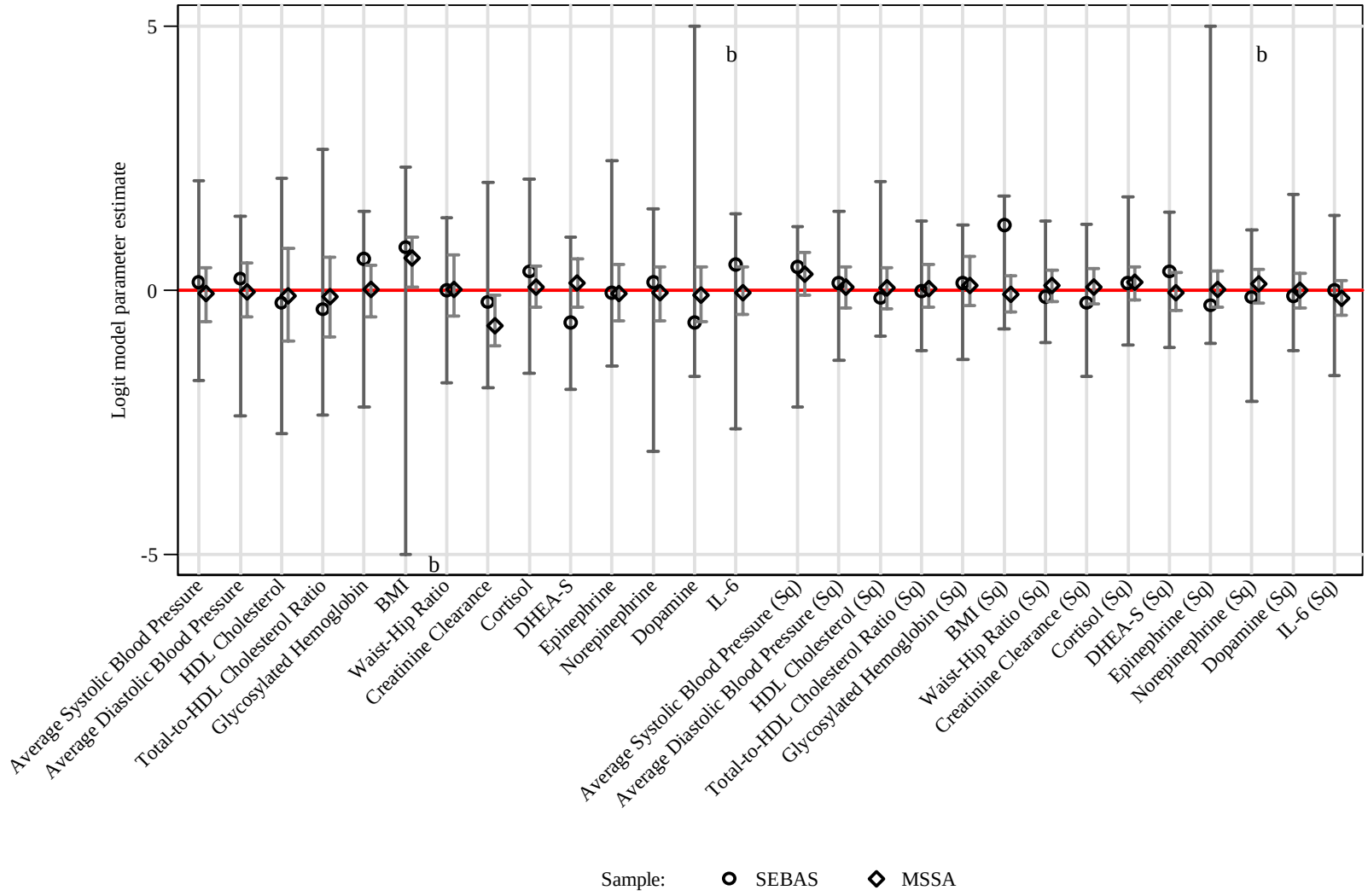
[a]Sample N = 139 for the SEBAS models and 259 for the MSSA models.
[b]Symbol key: ○ = SEBAS and ◊ = MSSA

Figure 4: Parameter estimates from separate logit models predicting at least one incorrect cognitive evaluation response at follow-up[a,b]



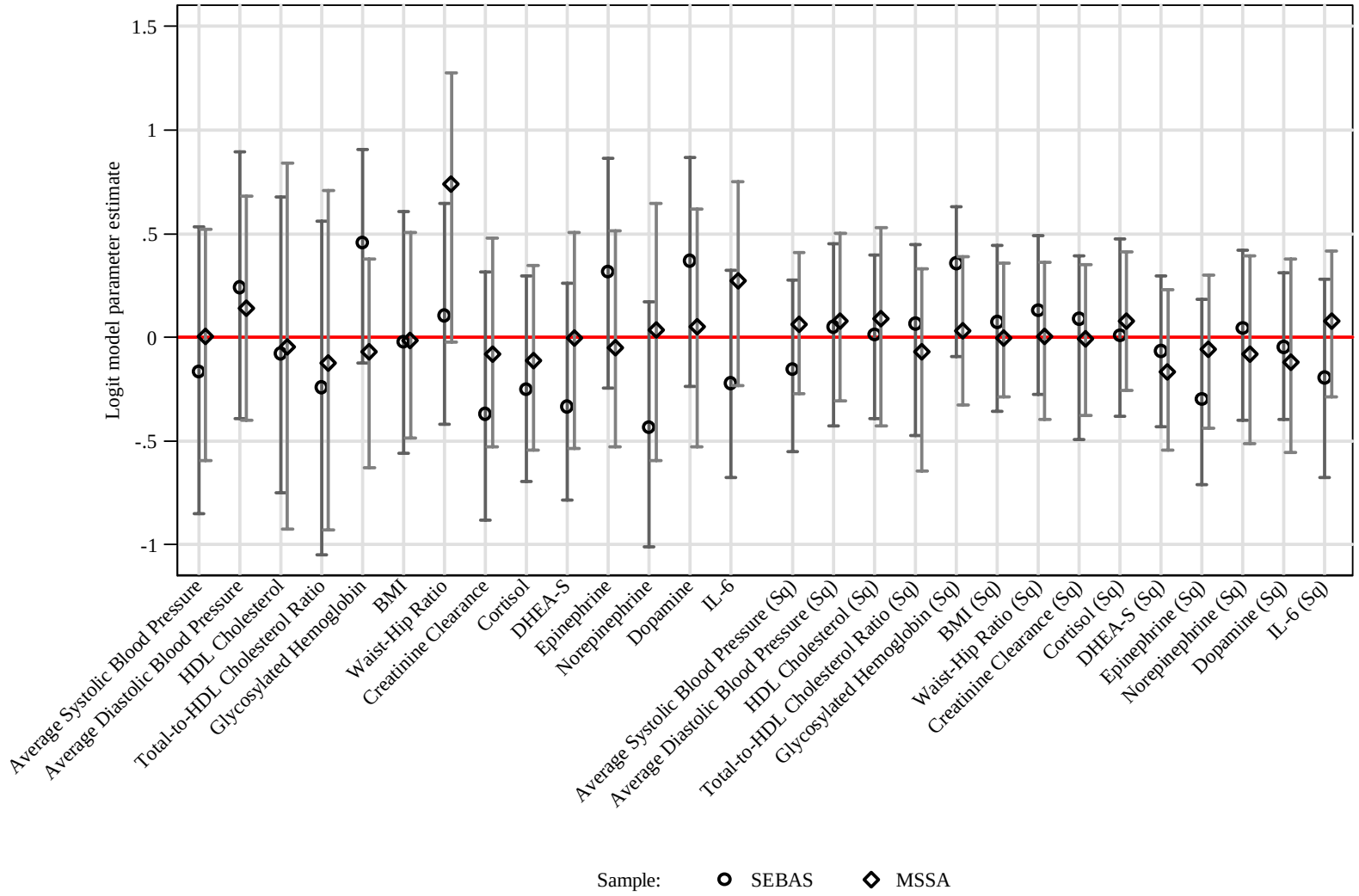
[a]Sample N = 235 for the SEBAS models and 215 for the MSSA models.
[b]Symbol key: ○ = SEBAS and ◊ = MSSA

Figure 5: Estimates from combined logit model predicting any mobility limitation at follow-up (bootstrapped)[a,b]



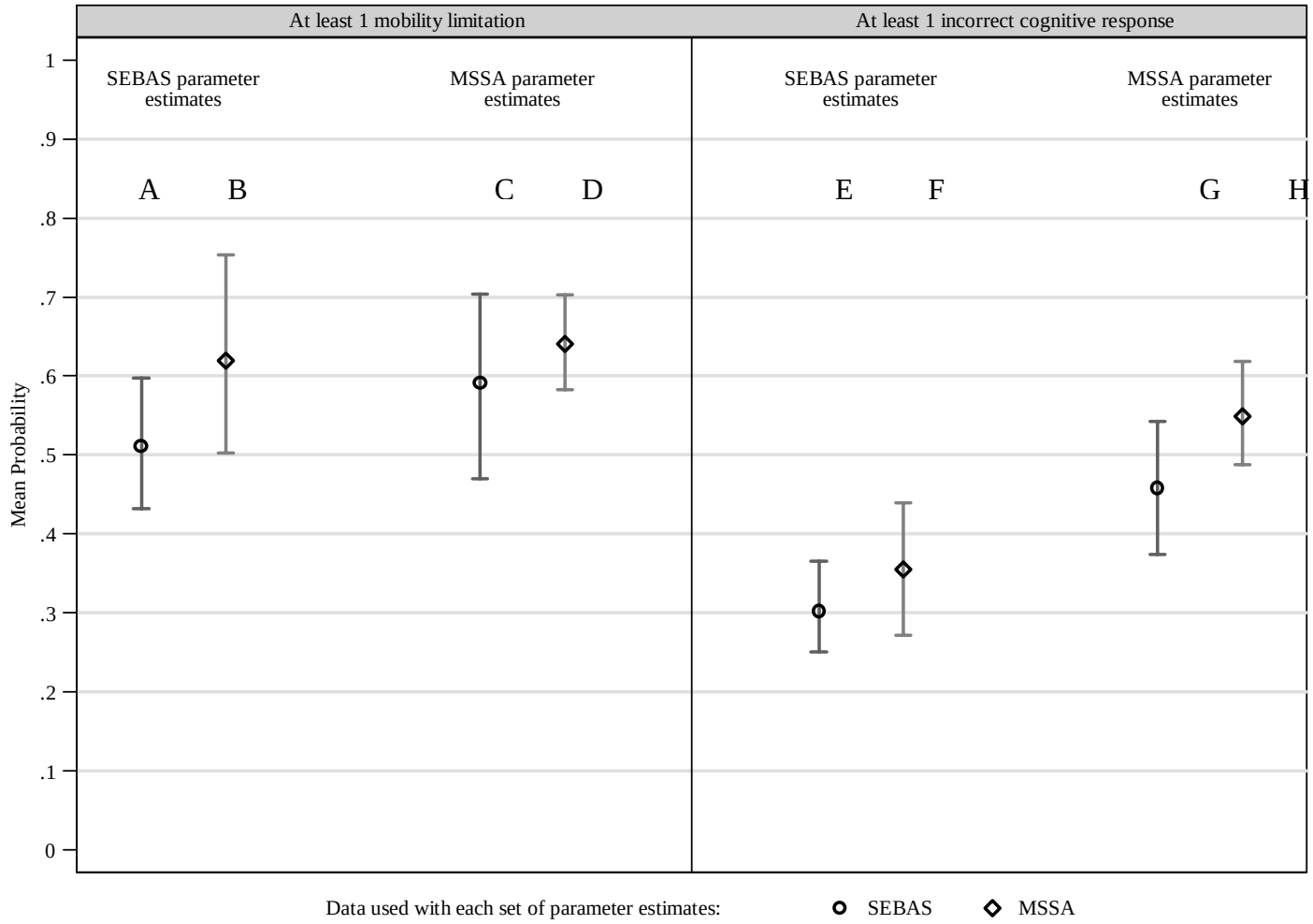
[a]Sample N= 139 for the SEBAS model and 259 for the MSSA model.
[b]Bias corrected bootstrapped 95% confidence intervals shown. In addition, calculations yielded very large estimates for three confidence interval parameters so, to facilitate graphing, these three values are capped to equal +/-5.

Figure 6: Estimates from combined logit model predicting at least one incorrect cognitive evaluation response at follow-up (bootstrapped)[a,b]



[a]Sample N= 235 for the SEBAS model and 215 for the MSSA model.
[b]Bias corrected bootstrapped 95% confidence intervals shown.

Figure 7: Bootstrapped mean predicted probabilities and 95% confidence intervals for each health outcome at follow-up, within and out-of-sample[a,b]



[a]Sample N= 139 for the SEBAS and 259 for the MSSA mobility models, and N= 235 for the SEBAS and 215 for the MSSA cognitive response models.

[b]Bias corrected bootstrapped 95% confidence intervals shown.