



Published in final edited form as:

*Eur J Epidemiol.* 2011 July ; 26(7): 511–525. doi:10.1007/s10654-011-9585-2.

## Reliability of hypothalamic–pituitary–adrenal axis assessment methods for use in population-based studies

**Sherita Hill Golden,**

Department of Medicine, Johns Hopkins University, Baltimore, MD, USA

Department of Epidemiology, Johns Hopkins University, Baltimore, MD, USA

Division of Endocrinology and Metabolism, Johns Hopkins University School of Medicine, 2024 E. Monument Street, Suite 2-600, Baltimore, MD 21287, USA

**Gary S. Wand,**

Department of Medicine, Johns Hopkins University, Baltimore, MD, USA

**Saurabh Malhotra,**

Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA

**Ihab Kamel,** and

Department of Radiology, Johns Hopkins University, Baltimore, MD, USA

**Karen Horton**

Department of Radiology, Johns Hopkins University, Baltimore, MD, USA

Sherita Hill Golden: sahill@jhmi.edu

### Abstract

Population-based studies have been hampered in exploring hypothalamic–pituitary–adrenal axis (HPA) activity as a potential explanatory link between stress-related and metabolic disorders due to their lack of incorporation of reliable measures of chronic cortisol exposure. The purpose of this review is to summarize current literature on the reliability of HPA axis measures and to discuss the feasibility of performing them in population-based studies. We identified articles through PubMed using search terms related to cortisol, HPA axis, adrenal imaging, and reliability. The diurnal salivary cortisol curve (generated from multiple salivary samples from awakening to midnight) and 11 p.m. salivary cortisol had the highest between-visit reliabilities ( $r = 0.63–0.84$  and  $0.78$ , respectively). The cortisol awakening response and dexamethasone-suppressed cortisol had the next highest between-visit reliabilities ( $r = 0.33–0.67$  and  $0.42–0.66$ , respectively). Based on our own data, the inter-reader reliability ( $r_s$ ) of adrenal gland volume from non-contrast CT was  $0.67–0.71$  for the left and  $0.47–0.70$  for the right adrenal glands. While a single 8 a.m. salivary cortisol is one of the easiest measures to perform, it had the lowest between-visit reliability ( $R = 0.18–0.47$ ). Based on the current literature, use of sampling multiple salivary cortisol measures across the diurnal curve (with awakening cortisol), dexamethasone-suppressed cortisol, and adrenal gland volume are measures of HPA axis tone with similar between-visit reliabilities which likely reflect chronic cortisol burden and are feasible to perform in population-based studies.

## Keywords

Adrenal gland volume; Cortisol awakening response; Cortisol diurnal curve; Dexamethasone suppression test; Reliability; Salivary cortisol

---

## Introduction

Allostasis is the central process employed by mammals to maintain homeostasis threatened by various forms of stress. It includes a series of dynamic actions through which a variety of neuroendocrine hormones, immune factors and autonomic nervous system mediators are triggered [1, 2]. When burdened by cumulative stress, the allostatic load (i.e., hypothetical measure of cumulative stress) of an organism increases, resulting in wear and tear on the organism from excessive exposure to the catabolic properties of glucocorticoids, stress peptides and pro-inflammatory cytokines. This burden taxes metabolic systems and can influence the development of insulin resistance, cardiovascular disease, osteoporosis and other disorders. One of the major contributors to allostatic load is cortisol exposure, regulated by the hypothalamic–pituitary–adrenal (HPA) axis.

## Physiology of the HPA axis

The HPA axis is a tightly regulated system that represents one of the body's response mechanisms to acute and chronic physiological or psychological stress. In response to physiological or psychological stressors (Fig. 1), the HPA axis is activated, resulting in release of CRH from the hypothalamus, which stimulates the anterior pituitary gland to release ACTH. ACTH stimulates release of cortisol from the adrenal glands, which results in a cascade of physiological events. Once the stressor has resolved the response is terminated through a negative feedback loop, in which cortisol suppresses further release of ACTH and CRH. Chronic stress injures this component of the stress response. Population-based studies often attempt to measure the contribution of the HPA axis to metabolic outcomes.

## Cortisol burden as a potential mediator between chronic stress and metabolic dysfunction in population-based studies

There are numerous environmental and genetic factors that can increase an individual's exposure to cortisol. For example, there is a growing body of evidence showing that clinical depression and depressive symptoms (states associated with hypercortisolism) predict development of cardiovascular disease [3] and type 2 diabetes [4]. While much of this association is explained by obesity-promoting health behaviors, these factors do not explain all of the observed associations [5]. There is also evidence accumulating that various forms of chronic stress promote the development of metabolic dysfunction [2]. It has been proposed that an increased allostatic load, primarily mediated by excess cortisol burden, may result in metabolic dysfunction documented in patients with depression and other conditions that chronically elevate cortisol exposure [6].

Population-based studies have been hampered in exploring a neuroendocrine link between these conditions due to lack of incorporation of reliable measures of chronic cortisol exposure which would permit quantifying the metabolic burden imposed by cortisol production. One major problem in selecting and interpreting cortisol measurements in epidemiological studies is that most existing measures reflect cortisol exposure over a short duration of time and may not reliably quantify the allostatic load imposed by chronic cortisol exposure over time. Another problem in selecting and interpreting cortisol measures is the need to be aware of the test-retest reliability of cortisol procedures, which vary widely, and only a few are useful for epidemiological studies. In addition, certain measurements of

HPA axis tone, such as overnight and 24-h urine free cortisol, laboratory-based stress tests and measurement of the hypothalamic and pituitary hormones—corticotrophin releasing hormone (CRH) and adrenocorticotrophic hormone (ACTH), respectively—are cumbersome to measure in population-based studies. In particular, CRH and ACTH are labile and require immediate laboratory processing to avoid sample degradation and they are pulsatile, such that a one-time measurement is unlikely to reflect diurnal activity, which is best obtained by frequent blood sampling every hour or less over 24-h. Hence, their measurement is not feasible in field settings.

The purpose of this review article is two-fold. First, we will summarize the current literature on the reliability of several measures of HPA axis tone, including one-time salivary cortisol measurement, cortisol awakening response, multiple salivary cortisol samples collected from awakening to bedtime, dexamethasone-suppressed cortisol, and adrenal gland volume. Second, we will discuss the feasibility and pros and cons of performing these measures in large, population-based studies. Measures of HPA axis responsivity (e.g., ACTH and CRH stimulation test, Trier psychosocial stress test, insulin-induced hypoglycemia), which may not assess chronic cortisol burden or are too invasive to use in non-clinical field settings, will not be discussed.

## Methods

### Search strategy

A search strategy was developed for MEDLINE using PubMed, with a combination of controlled vocabulary (MeSH terms) and key word terms and phrases to depict the concept of cortisol, HPA axis, adrenal imaging, and reliability (see Appendix). We limited the strategy to Human and English-language articles published through June 2010, and excluded review articles. Search terms for salivary cortisol were included because it is currently used in population-based studies and considered to be a feasible and non-invasive measure of cortisol. We identified 3,516 articles and reviewed the titles and abstracts. Our primary criteria for article inclusion were studies that included reliability data (see below) on repeated measures of HPA axis separated by at least 24 h. Articles were excluded if they did not compare repeated measures of HPA axis function, utilized stimulation testing other than the dexamethasone-suppression test (i.e. CRH and ACTH testing, mental stress testing), were treatment studies of hypo- or hypercortisolism, assessed brain imaging, or were not relevant to the objectives of the review ( $n = 3,497$ ). We also excluded articles that focused on brain imaging of the pituitary gland and other structures because brain imaging is expensive and requires significant radiation exposure for population-based studies. We did, however, include articles that assessed adrenal gland volume even though its assessment is also expensive and accompanied by radiation exposure. Because many population-based studies perform body CT and MRI scans to measure coronary artery calcium and intra-abdominal fat, assessment of adrenal gland volume can be feasibly assessed simultaneously without significant additional participant burden in field settings. We identified 19 articles that assessed repeated measures of one-time salivary cortisol measurement, cortisol awakening response, multiple salivary cortisol samples collected from awakening to bedtime, dexamethasone-suppressed cortisol, and adrenal gland volume. These articles are summarized in the results section of this review.

### Measures of reliability of laboratory measurements

Assessing the reliability of a biological measure determines the extent to which results agree when measured using the same approach at different time points or when using different approaches (i.e. different observers). In epidemiological studies, we are attempting to measure the variability between study participants; however, there can also be variability

within study participants due to (1) variability of the laboratory measurement method or (2) variability in participant behaviors and physiology [7]. Intra-observer variability is due to variability in a given laboratory measurement performed on the same participant at different points in time. Inter-observer variability is due to variability in a given laboratory measurement conducted on the same sample by different technicians [7]. In this review, we will be focused primarily on intra-observer variability/reliability. There are several methods of assessing test-retest reliability of continuous biological measures. Most studies examining the reliability of HPA axis measures have used the intraclass (ICC or R), Pearson's linear ( $r$ ), or Spearman's ordinal ( $r_s$ ) correlation coefficient [7]. ICC is a true measure of agreement that combines information on the correlation and systematic differences between readings [7]. While the Pearson's  $r$  is often used as the main measure of reliability, it is sensitive to the range of values as well as outlier [7]. Spearman's  $r$  is less influenced by outliers but neither measure accounts for systematic biases between repeated measures [7]. Throughout the review, we will indicate which reliability measure was used by each study summarized.

## Results

### Salivary cortisol

Measurement of salivary cortisol has the advantage of being easy to perform in large studies in the free-living state. It has several other advantages, including being (1) non-invasive, (2) amenable to timed sample collections in the free-living state without the need for medical personnel, (3) being stable at room temperature for a least 1 week and thus can be mailed back to the investigator, (4) measuring free or the physiologically active form of cortisol, and (5) having a strong correlation to free cortisol measured in plasma and serum [8–11]. Salivary cortisol levels are also stable following repeated cycles of freezing and thawing and cortisol levels in centrifuged saliva samples may be stored at 5°C for 3 months or –20 and –80°C for a least 1 year [12]. Because cortisol measured in saliva is free and not bound, it is not as subject to variation by factors that affect cortisol binding globulin (CBG), the primary transport protein for serum cortisol. Concentration of plasma CBG is altered by oral contraceptive use, pregnancy, severe illness and liver disease. Thus serum cortisol measurements can be misleading under these conditions. Although it is possible to measure unbound cortisol in serum, the test is expensive and, unlike salivary cortisol, requires venipuncture [13].

These virtues have led investigators to consider collection of salivary cortisol samples to characterize HPA axis tone. However, there are several limitations to consider in using salivary cortisol measurements to assess HPA axis tone, including (1) contamination of the salivary cortisol collection device by over-the-counter hydrocortisone creams and ointments, salivary blood, or consumption of low pH substances (which can artificially raise cortisol levels), (2) non-compliance with the recommended sample collection time, (3) insufficient saliva collection, and (4) the effect of smoking (current smoking is associated with higher levels than non- and former smoking) [10, 13–15]. Additionally, several medications (e.g. anti-depressants; oral, nasal, topical, and ophthalmic corticosteroids; alpha, beta, and cholinergic receptor antagonists) have the potential to impact salivary cortisol levels; however, few behaviorally oriented studies have comprehensively documented medication use or focused on the impact of various medication classes on salivary cortisol levels [16].

Despite these limitations, salivary cortisol still represents a practical approach to assessing the HPA axis in large, population-based studies in which collection of repeated blood samples is impractical and/or infeasible. The issue that is often faced in designing epidemiological studies is how many salivary cortisol samples should be collected to properly characterize HPA axis tone. Cortisol is a pulsatile hormone and has large fluctuations in blood reflecting both stressed and non-stressed states. Does a single salivary

measurement per 24 h period reflect cortisol exposure during the 24 h period or does it measure cortisol exposure at only that moment in time of collection? Does acquiring multiple samples throughout the day reflect cortisol burden over days, weeks or months or does it merely reflect cortisol burden limited to the 24 h period of sampling? This conundrum is not limited to salivary cortisol but is an unanswered question for every measurement of HPA axis tone. Indeed, while using a single measurement is appealing, prior studies have suggested that a single salivary measurement does not have adequate between visit reliability, which would limit its use in longitudinal studies attempting to characterize an individual's chronic HPA axis tone.

### Reliability of one time 8 a.m. and 11 p.m. salivary cortisol

As with many other hormones, cortisol has a circadian variation, characterized by peaks in the early morning hours (8 a.m.) and nadirs in the late evening [8, 13]. Often an 8 a.m. or late night (i.e. 11 p.m. or midnight) salivary cortisol sampling will be collected under the assumption that individuals with high cortisol burden will have elevated cortisol levels at both the peak and nadir of the cortisol circadian rhythm. Thus such measurements may represent options for assessing chronic HPA axis tone in population-based studies. Table 1 summarizes 3 studies that have examined the between-visit reliability of a single 8 a.m. salivary cortisol. Among 20 healthy males, the between visit reliability (R) of 8 a.m. salivary cortisol collected 1–5 weeks apart was 0.18 [17]. In another study of 116 women without Major Depressive Disorder, the between visit reliability ( $r_s$ ) of 8 a.m. salivary cortisol samples collected 6 months apart was 0.41 ( $P < 0.001$ ) [18]. Similarly, the between visit reliability (R) for repeated measures of morning salivary cortisol samples in 20 healthy volunteers was 0.47, indicating poor reproducibility with 53% of the variance explained by intra-individual differences [19]. We collected quality control data on the between-visit repeatability of one time 8 a.m. salivary cortisol on a subset of 146 male and female participants in the Atherosclerosis Risk In Communities (ARIC) Carotid MRI Study (see Table 2). Among 48 repeat visit quality control replicates with sampling separated by an average of 81.7 days, the intraclass correlation coefficient (R) was quite low, 0.27, suggesting that one time morning salivary cortisol sampling is not an appropriate surrogate marker which reflects chronic cortisol burden over weeks to months. These findings were similar to those of Coste et al.

Because the morning cortisol may be affected by several extrinsic factors related to awakening, 11 p.m. salivary cortisol might be an additional alternative for a one-time salivary cortisol collection. Based on the normal diurnal secretion of cortisol, it should reach a nadir between 2300 and 0000 [13]. In the same study of 20 healthy volunteers discussed above, the intraclass correlation coefficient (R) for repeated measures of 11 p.m. salivary cortisol was 0.78, indicating that 22% of the variance was explained by intra-individual differences in measurements [19]. While the reliability of the 11 p.m. salivary cortisol was better in this population than the 8 a.m. salivary cortisol, the duration of time between repeated sample measurements was likely days, so it is unclear if similar reproducibility of results would be found on sample collections separated by weeks to months. Finally, in a study of individuals with subclinical hypercortisolism (as might be seen in individuals with depressive and metabolic disorders), the 11 p.m. salivary cortisol had a lower sensitivity in detecting subclinical hypercortisolism compared to the dexamethasone-suppressed cortisol (see below) or 24-h UFC [20]. This suggests that it might also be less sensitive in detecting subtle differences in subclinical hypercortisolism in a relatively healthy population, such as individuals enrolled in longitudinal, epidemiological studies. Thus we do not recommend the use of single samplings.

## Reliability of multiple salivary cortisol measurements from awakening to midnight

An alternative approach to collecting one-time salivary cortisol samples is to have participants collect several samples over the course of 12 h to measure cortisol exposure over a longer time period. An advantage to this approach is that it incorporates the awakening cortisol response (see below) as well as cortisol secretion throughout the day.

**Cortisol awakening response**—It is established that there is a pronounced cortisol awakening response (CAR). Both salivary and serum cortisol increases by 50–70% during the first 30 min after awakening and remains elevated for about 60 min [21]. The CAR is thought to reflect the adrenal capacity to respond to stress and can be exploited to capture subtle differences in HPA axis tone as a function of exposure to chronic stress. While affected by other factors, such as sex (women have a greater awakening response than men) and oral contraceptive use (users have a smaller awakening response than non-users), the salivary cortisol response to awakening has a higher intra-individual stability than a single morning salivary cortisol or measurement of salivary cortisol at predefined times [21].

A prior study summarized data on the awakening salivary cortisol response on three populations of individuals—42 children (mean age  $11.2 \pm 2$  years), 70 young adults (mean age  $26.5 \pm 6.3$  years), and 40 elderly individuals (mean age  $70.4 \pm 5.7$  years) [21]. In the children, where the CAR area under the curve (AUC) was characterized by samples collected 0, 10, 20, and 30 min after awakening, the intra-individual correlation ( $r$ ) between repeated measures of the CAR AUC separated by a day over 3 days ranged from 0.39 to 0.67 (all  $P < 0.05$ ). In the elderly, where the CAR AUC was characterized by samples collected 0, 15, 30, and 60 min after awakening, the intra-individual correlation ( $r$ ) for CAR AUC separated by a day was 0.58 ( $P < 0.05$ ) [21]. In the young adults, awakening salivary cortisol samples were collected at 0, 15, 30, and 60 min after awakening on 3 occasions separated by 1 week and the intra-individual correlation ( $r$ ) ranged from 0.42 to 0.65 (all  $P < 0.05$ ) [21]. These data suggest that the 16–45% of the variance in the CAR AUC was explained by intra-individual differences, indicating moderate to high stability of the AUC cortisol levels across days to weeks [21]. Similar findings were observed in a population of 42 healthy volunteers (76% women, mean age 35 years), where the CAR AUC was characterized by samples collected at 0, 15, 30, and 45 min after awakening on two consecutive days. In this study, the intra-individual correlation ( $r$ ) between repeated measures of the salivary CAR AUC was 0.34–0.52 (all  $P < 0.05$ ) [22].

For the CAR to be informative, it is important that subjects collect samples at the specified times following awakening. Electronic monitoring devices on the salivary cortisol collection devices should be used detect deviations from the protocol and maximize compliance and accuracy in the documentation of sample times.

### Multiple daily salivary cortisol measurements from awakening to midnight—

Another multiple cortisol measurement technique is to measure cortisol through all or part of the 24 h circadian cycle. Table 1 summarizes 3 studies that have examined the between-visit reliability of measuring salivary cortisol in this manner. In the study [22] summarized above, in addition to collecting awakening salivary cortisol samples, the authors also collected 4 additional samples 3, 6, 9, and 12 h after awakening on two consecutive days. For this analysis the CAR was excluded and the curve was characterized using the 0, 3, 6, 9, and 12 h post-awakening samples. The intra-individual correlation ( $r$ ) between repeated measures of the diurnal cortisol curve separated by one day was 0.647 ( $P < 0.0001$ ) [22]. Another study assessed the diurnal cortisol profile in a population of 50 older adults by collecting salivary cortisol samples at awakening, 30 min post-awakening, 5 p.m. and 9 p.m. on two consecutive days [23]. They found that when the awakening sample was used as the anchor

(as opposed to the wake + 30 min post-awakening sample), the between-visit reliability ( $r_s$ ) of the diurnal cortisol slope was 0.63 ( $P < 0.05$ ) [23]. The predicted test-retest reliability ( $r_s$ ) of samples collected over 2 and 3 days was 0.78 and 0.84, respectively, suggesting that at least 2 days of sample collection are necessary to adequately characterize the diurnal curve [23]. These studies were limited to being conducted in small populations.

While collecting multiple timed saliva samples might be cumbersome to perform in a population-based study, it has already been successfully performed in 1,000 participants in the Multi-Ethnic Study of Atherosclerosis (MESA) [24]. A potential disadvantage to having study participants collect several salivary cortisol samples, particularly if they are timed, is that compliance might be reduced [14]; however, in the MESA Study, participants were compliant with 6 salivary cortisol sample collections/day over 3 consecutive weekdays [24]. In MESA, salivary cortisol samples were collected in 936 men and women at the following times—directly upon waking; 30 min after waking; 10:00 a.m.; 12:00 p.m. or before lunch, whichever was earlier; 6:00 p.m. or before dinner, whichever was earlier; and at bedtime. This daily collection protocol was repeated on each of three successive weekdays. The intra-individual correlation ( $r$ ) between mean salivary values across diurnal curves over the 3 days ranged from 0.65 to 0.72 [24]. Another advantage to collecting multiple salivary cortisol samples from awakening to bedtime is that multilevel, mixed, or hierarchical linear statistical models can be used to simultaneously account for within and between individual differences in cortisol measures [25].

Very few studies have evaluated the between-visit repeatability of integrated daily salivary cortisol whose measurements are separated by a longer time interval. A small study of 28 children found reproducibility in repeated cortisol measurements at different time points throughout the day as well as in cortisol area under the curve in measurements separated by a median interval of 1 year; however, the specific correlation between the repeated measures was not reported [26]. Further studies are needed to determine the repeatability of integrated salivary cortisol over longer time intervals (i.e. weeks to months).

Is it possible to derive useful information collecting fewer cortisol measurements? Kraemer et al. suggests that 2 salivary cortisol samples collected at awakening and 9 p.m. may adequately characterize the diurnal cortisol slope [23]. In their study of 50 adults, the slope of the awakening to 9 p.m. cortisol correlated strongly with the slope determined from all 5 sample collections ( $r_s = 0.954$ ; 95% CI: 0.530–0.974) [23]. When the cortisol slope was characterized by fewer than 5 samples, those calculated including the awakening and 9 p.m. salivary cortisols (regardless of the others chosen) were more strongly correlated to the slope calculated from all samples ( $r_s = 0.957$ – $0.976$ ; all  $P < 0.05$ ) than those that did not include the 9 p.m. sample ( $r_s = 0.281$ – $0.601$ ;  $P$ -values non-significant to  $P < 0.05$ ) [23]. These findings need to be confirmed in a larger, population-based sample of men and women of various ethnicities.

**Relation of CAR to multiple daily salivary cortisol measurements from awakening to midnight**—A prior study of 22 healthy adults failed to find an association between the CAR and integrated 12-h salivary cortisol measured every 15 min from 9 a.m. to 9 p.m. [27]. A more recent study by Edwards et al., however, suggests the CAR correlates with total daily cortisol secretion. In their study of 42 healthy men and women, there was a significant positive association between the CAR area under the curve and 12 h mean diurnal cortisol on two separate days ( $r = 0.595$  and  $r = 0.660$ ; both  $P < 0.0001$ ), indicating that individuals who secreted more cortisol within the first 45 min after awakening secreted more cortisol throughout the day [22]. Their data suggest that the CAR may represent a measure that predicts cortisol levels throughout the remainder of the day; however, additional data from larger studies are needed to confirm this observation.

## Dexamethasone-suppressed cortisol

The dexamethasone suppression test is widely used in clinical endocrinology during the investigation of Cushing syndrome. Individuals with clinical and/or subclinical hypercortisolism fail to adequately suppress serum cortisol levels at 8 a.m. in response to 1 mg of oral dexamethasone administered at 11 p.m. the night before [28]. In the research setting, doses of dexamethasone below 1 mg (e.g. 0.5 or 0.25 mg) have been shown to allow detection of subtle degrees of increased HPA axis tone [29, 30], as might be present in community-dwelling study participants in a non-clinical cohort. In both non-clinical and clinical cohorts, the reduction in negative feedback sensitivity is thought to reflect the influences of chronic stress on the HPA axis. We have previously shown, in 20 healthy African-American women (mean age  $32 \pm 8$  years) without affective illness or any form of Cushing's or Pseudo-Cushing's Syndrome, that 8 a.m. salivary cortisol levels following administration of 0.5 mg of dexamethasone correlated strongly with daily salivary cortisol measurements collected at several time points after awakening (0800, 0845, 1030, 1600, 2000, and 2300;  $r_s = 0.44-0.77$ ; all  $P < 0.05$ ) [31]. Inadequate suppression of cortisol following dexamethasone administration indicates injury to the negative feedback mechanisms; this defect increases systemic cortisol exposure resulting in greater cortisol burden. These data suggest that the dexamethasone suppression test and multiple measurements of salivary cortisol throughout the day may provide similar information regarding cortisol burden, although our findings should be confirmed in a larger cohort.

Even though the dexamethasone suppression test is a stimulation test, it is more feasible to perform in population-based studies than other stimulation tests (e.g. CRH and ACTH stimulation tests) because (1) the medication (dexamethasone) can be taken orally at home by the participant, as opposed to the IV administration of ACTH and CRH, and (2) it only involves one sample collection which can be blood or saliva (see below).

**Reliability of the dexamethasone-suppressed cortisol**—The dexamethasone-suppressed cortisol is stable over time, even with changing metabolic conditions (Table 1). Yanovski et al. studied the effects of weight loss on dexamethasone-suppressed cortisol in obese binge and nonbinge eaters who lost 16–17% of their body weight and found that this HPA axis measure was not significantly changed from baseline to the 12-weeks return visit [32]. Thus, dexamethasone-suppressed cortisol was stable in the setting of an intervention that did not specifically target HPA axis tone. Even though personality traits and affective states are known to affect HPA axis tone, in a small study of 13 women with Borderline Personality Disorder with or without Post-Traumatic Stress Disorder, the cortisol response to dexamethasone suppression was stable over 1 year. There was also a correlation between 8 a.m. morning ( $r = 0.42$ ;  $P = 0.076$ ) and 4 p.m. evening ( $r = 0.56$ ;  $P = 0.023$ ) cortisol following dexamethasone suppression [33]. In a large study which examined the repeatability of dexamethasone-suppressed cortisol in 164 healthy elderly individuals without endocrine or affective disorders, there was inter-person variability but intra-person stability of the response to the dexamethasone suppression test over 2.5 years [30]. Inter-person variability is thought to reflect differences in cortisol burden across individuals. In this study, there was a strong correlation between plasma cortisol following two dexamethasone suppression tests separated by 2.5 years, which was not affected by age ( $r_s = 0.66$ ;  $P < 0.001$ ) [30].

**Use of the dexamethasone suppression test with salivary versus serum cortisol**—While disadvantages to using this test are the requirement for medication administration and a timed collection of serum cortisol, an advantage to its use in the field setting is that it can be performed with collection of salivary cortisol, negating the need for a blood draw. While one small study in 29 individuals found that salivary cortisol was more



variable and less repeatable than serum cortisol following dexamethasone administration [34], they and others found that the fractional suppression of serum and salivary cortisol in response to dexamethasone were similar [34–36]. In addition, there was a strong correlation between post-dexamethasone serum and salivary cortisol measurements [35, 36]. In a large study of 250 psychiatric inpatients undergoing a 1 mg dexamethasone suppression test, the correlation ( $r$ ) between post-dexamethasone serum and salivary cortisol was 0.89 [36]. Therefore, measuring salivary instead of serum cortisol is likely justifiable to minimize participant burden and reduce costs large population-based studies.

### Adrenal gland volume

Chronic activation of the HPA axis increases adrenal gland volume due to the trophic effects of ACTH on the adrenal cortex. Adrenal gland volume is thought to be a stable measure over several weeks to months in a stable environment but is also responsive to changing clinical conditions. In individuals with proven ACTH-dependent Cushing's syndrome, there is a significant correlation between adrenal gland width and plasma cortisol and 24-h urine free cortisol [37]. There is also a significant correlation between estimated disease duration and adrenal gland width, suggesting that it is an index of chronic HPA axis tone and glucocorticoid exposure over time [37]. Major Depressive Disorder, often accompanied by persistent activation of the HPA axis, is associated with adrenal hypertrophy, which resolves over several weeks to months following remission of depression and resolution of HPA axis hyperactivity [38]. The dose dependent nature of this effect is evidenced by the development of adrenal atrophy observed following injury to ACTH secretion as seen in individuals with central adrenal insufficiency [39, 40].

Adrenal gland volume may thus represent an integrated, non-invasive measure of HPA axis tone. We previously showed among healthy African-American women, adrenal gland volume was strongly correlated with cortisol following dexamethasone suppression ( $r_s = 0.66$ ;  $P = 0.004$ ), suggesting that women with higher cortisol following dexamethasone administration had increased adrenal gland volume and higher cortisol burden [31]. Adrenal gland volume did not correlate with 24-h urine free cortisol, as shown in prior studies [41, 42].

**Assessment of adrenal gland volume**—Adrenal gland volume is assessed using CT or magnetic resonance imaging (MRI). An advantage to using either of these two radiological approaches in population-based studies is that the adrenal volume can be assessed in individuals already receiving scans for other reasons as part of the parent study. For example, since many population-based studies perform body CT scans to measure coronary artery calcium and intra-abdominal fat, assessment of adrenal gland volume can be feasibly assessed simultaneously without significant additional participant burden. The additional scan time for performing a thin-slice adrenal protocol as part of a pre-existing cardiac or abdominal CT scan is only 5–10 min. The actual adrenal volume measurements, however, described below, are labor-intensive to perform and require 30–45 min/scan committed by a radiologist or highly-skilled technician. A disadvantage to using CT is that it requires radiation exposure (approximately 320 mrem, comparable to annual background radiation exposure). Both CT and MRI are expensive.

**Reliability of adrenal gland volume**—There is a paucity of information on intra-individual repeatability or inter-reader reliability of adrenal gland volume. In a small study of 22 pheochromocytoma patients who underwent subtotal adrenalectomy, an adrenal CT scan with intravenous contrast was performed on postoperative day 4, and repeated 3 months later to assess the volume of the adrenal remnant [43]. The between-visit repeatability ( $r$ ) of the adrenal remnant volume was 0.84 ( $P < 0.05$ ) [43]. Very recently, a small pilot study of 4

individuals sought to determine the intra- and inter-observer variation and repeatability of adrenal volume measurements obtained from MRI. Adrenal MRIs were separated by 7 days in each participant. They found that between-visit variation in the adrenal volumes was small (5% of a 3-cm<sup>3</sup> adrenal gland) and while the intra- and inter-observer variation were larger, they were similar and still small (9% of a 3 cm<sup>3</sup> adrenal gland) [44].

A disadvantage to using contrast for adrenal imaging by CT is that it precludes measurement in individuals with renal disease or dye-induced allergies. Non-contrast CT is an alternative. We did not identify published data on intra-individual reliability or inter-reader reliability of adrenal gland volume from non-contrast CT; however, we set out to generate inter-reader reliability data based on data from our population of healthy women. As previously described [31], we performed abdominal CT scans to measure adrenal gland volume in 28 healthy African-American women ages 18–45 years using a Siemens Multidetector-Row Scanner (either 16 or 64 slice) [31]. One hundred and twenty 1 mm slices were made through the adrenals and the contour was manually traced on each slice with a console cursor. The GE Advantage Workstation software, which is commercially available, was used to automatically calculate the adrenal volume by summing the area on each slice [31].

Figure 2 shows the anatomic location of the left and right adrenal glands. We hypothesized that inter-reader reliability of the left adrenal gland might be more reproducible because the right adrenal gland is harder to distinguish from surrounding structures and is often compressed against the spine between the liver and inferior vena cava, as shown in Fig. 2. We also hypothesized that women with higher body-mass index (BMI) might have more retroperitoneal fat, making the adrenal glands easier to distinguish from surrounding tissue in heavier women. We therefore, determined the inter-reader reliability of the left and right adrenal glands separately, stratified by BMI, as summarized in Table 3.

Two radiologists performed adrenal volume calculations on 28 study participants. Among overweight and obese women (BMI  $\geq 25$  kg/m<sup>2</sup>) the inter-reader reliability determined by two independent readers separated by 12–24 months was very good for the left adrenal gland volume ( $r_s = 0.71$ ;  $P = 0.001$ ) but not as good for the right adrenal gland volume ( $r_s = 0.47$ ;  $P = 0.056$ ). Among lean women with BMI  $< 25$  kg/m<sup>2</sup>, the inter-reader reliability of left and right adrenal gland volumes were similar at  $r_s = 0.67$  ( $P = 0.02$ ) and  $r_s = 0.70$  ( $P = 0.02$ ), respectively. While we found similar inter-reader reliability of the left and right adrenal gland volumes in lean women, we found a discrepancy in the overweight and obese women. In addition, the small size of the adrenal glands requires significant technical skill to accurately determine volumes on multiple CT slices.

## Discussion

Despite the challenges associated with measuring chronic HPA axis tone, it is important to identify reliable measures for incorporation into population-based cohort studies to determine if they provide an additional explanatory mechanism for the association between chronic stress (i.e. neighborhood violence, racism, poverty) and enhanced metabolic risk [45]. In order to expand the field of epidemiology to include assessment of HPA axis tone, it is important to recognize the balance between identifying reliable hormonal measures and identifying measures that are feasible to incorporate into large studies in the field setting.

### Reliability and feasibility of salivary cortisol-associated HPA axis measures

In considering several measures of chronic HPA axis tone reviewed in this article, the total diurnal salivary cortisol curve and 11 p.m. salivary cortisol measurement have the highest between-visit reliability ( $r = 0.63$ – $0.84$  and  $0.78$ , respectively). While the 11 p.m. salivary cortisol has the advantage of being easy to perform in the free-living state, collecting

multiple daily samples over 2–3 days to characterize the diurnal salivary cortisol curve is more cumbersome and imposes greater participant burden. Despite these limitations, the latter has been successfully performed in the MESA Study [24]. Future studies need to confirm whether the between visit reliabilities are similar for 11 p.m. salivary cortisol and the salivary cortisol curve generated from multiple daily salivary samples collected from awakening to midnight when visits are separated by weeks to months, as opposed to consecutive days.

Following the 11 p.m. salivary cortisol and the salivary cortisol curve generated from multiple daily salivary sample collections from awakening to midnight, the CAR and dexamethasone-suppressed cortisol had the next highest between-visit reliabilities ( $r = 0.33$ – $0.67$  and  $0.42$ – $0.66$ , respectively). The cortisol awakening response has a higher intra-individual reliability than a single 8 a.m. morning cortisol (see below); however, it is more cumbersome to perform in the population-based setting because participants require additional instruction about sample collection and accurate timing of sample collection in relation to awakening. Between visit reliability is stable over days to 3 weeks; however, additional studies are needed to determine whether the reliability is similar over many months. The dexamethasone-suppressed serum cortisol has the advantage of being one of the few measures of HPA axis tone which shows moderate intra-individual repeatability over 1–2 years, which is a longer timeframe than reliability data for other HPA axis measures. Prior studies of depression that have incorporated a dexamethasone suppression test show normalization of post-dexamethasone cortisol following clinical recovery [46, 47]. This indicates that the test can be modified as cortisol burden changes thus reflecting its usefulness as a measure to capture changes in allostatic load. We recognize that the dexamethasone-suppression/corticotrophin releasing hormone test is used in psychiatry to detect HPA axis hyperactivity and monitor response to therapy in depressive disorders [48, 49]; however, this test of HPA axis tone would add undue cost and participant burden to large epidemiological studies and is therefore, not a practical consideration in a non-clinical setting.

While a single 8 a.m. salivary cortisol is one of the easiest measures to perform, it generally has the lowest between-visit reliability ( $R = 0.18$ – $0.47$ ), whether sample collection was separated by days, weeks, or months. Thus, this likely represents the poorest measure of chronic HPA axis tone.

### Reliability and feasibility of adrenal gland volume

Based on our own data, the inter-reader reliability ( $r_s$ ) of adrenal gland volume from non-contrast CT ranged from 0.67 to 0.71 for the left and 0.47 to 0.70 for the right adrenal gland. Adrenal volume can be more easily determined with greater reliability on a contrast CT [43]; however, contrast administration may be complicated by allergic reactions and/or renal side effects. While MRI may represent an alternative assessment method that does not require radiation exposure, both CT and MRI scans are expensive to perform if not already a component of the main study. Finally, the reliability of adrenal volume measurements may be affected by inter-device differences if more than one scanner is used during a study. Despite these limitations, the inter-reader reliability of adrenal gland volume was similar to that of the salivary cortisol-based HPA axis measures.

### Conclusion

Based on the current literature, use of sampling multiple salivary cortisol measures across the diurnal curve (with the CAR), dexamethasone-suppressed salivary cortisol, and adrenal gland volume are measures of HPA axis tone with similar between-visit reliabilities. It is notable that HPA axis measures are generally not as reproducible as other biological

measures collected in observational studies (e.g. serum creatinine), which likely reflects the many biological and environmental factors that impact the HPA axis and free cortisol availability [9, 16]. Thus a note of caution is in order. Although these techniques capture cortisol burden over a certain time frame, additional studies are required to quantify this time period and to determine if the measures are surrogates for the cortisol burden that has been present for weeks or months.

A final consideration for use of HPA axis measures in population-based studies is participant burden. In larger, multi-site cohort studies in the United States, use of the dexamethasone suppression test, repeated collections of salivary cortisol, and radiological procedures (e.g. CT, MRI) requires review by the study Steering Committee as well a National Institutes of Health-appointed Observational Study Monitoring Board to assess participant burden and safety risks. However, it is feasible to incorporate these measures into the workflow of an on-going study exam. In order to advance our understanding of the biological relation between depression and stress-related disorders and metabolic outcomes, epidemiologists, endocrinologists, behavioral scientists, and laboratory scientists will need to work collaboratively in population-based research to continue to identify additional cortisol biomarkers and incorporate them into on-going studies.

## Acknowledgments

This review and accompanying studies were supported by the National Institute of Diabetes, Digestive, and Kidney Diseases, National Institutes of Health (K23 DK071565 to SHG) and by the National Institute of Alcohol Abuse and Alcoholism (RO1 AA10158 to GW).

## Abbreviations

<b>ACTH</b>	Adrenocorticotrophic hormone
<b>ARIC</b>	Atherosclerosis Risk In Communities Study
<b>AUC</b>	Area under the curve
<b>BMI</b>	Body mass index
<b>CAR</b>	Cortisol awakening response
<b>CBG</b>	Corticotrophin binding globulin
<b>CRH</b>	Corticotrophin releasing hormone
<b>CT</b>	Computed tomography
<b>HPA</b>	Hypothalamic-pituitary-adrenal axis
<b>ICC (R)</b>	Intraclass correlation coefficient
<b>MESA</b>	Multi-Ethnic Study of Atherosclerosis
<b>MRI</b>	Magnetic resonance imaging
<b>R</b>	Pearson's linear correlation coefficient
<b>r<sub>s</sub></b>	Spearman's ordinal correlation coefficient
<b>UFC</b>	Urine free cortisol

## Appendix: medline search strategy

“Salivary cortisol”[tiab] OR “Daytime cortisol”[tiab] OR “morning cortisol”[tiab] OR “Awakening cortisol”[tiab] OR “Free cortisol”[tiab] OR “diurnal cortisol”[tiab] OR

“cortisol”[ti] OR “cortisol response”[ti] OR “Cortisol levels”[tiab] OR “Cortisol rhythms”[tiab] OR “Cortisol rhythm”[tiab] OR “Cortisol secretion”[tiab] OR “Dexamethasone”[ti] OR “hydrocortisone”[tiab] OR “adreno-corticotrophic hormone”[tiab] OR “Saliva/chemistry”[majr] OR “Saliva/drug effects”[majr] OR “adrenocortical stress capacity”[All Fields] OR “magnetic resonance imaging”[tiab] AND (“Adrenal Gland Volume”[tiab] OR “adrenal insufficiency”[tiab] OR “Adrenocortical Insufficiency”[tiab] OR “Adrenocorticotrophic hormone deficiency”[All] OR “ACTH deficiency”[tiab] OR “Adrenocortical Hyperplasia”[tiab] OR “adrenocortical activity”[tiab] OR “adrenal incidentaloma”[tiab] OR “Adrenalectomy”[tiab] OR “Hypercortisolism”[tiab] OR “Cushing’s syndrome”[tiab] OR “Hypothalamic Pituitary Adrenal Axis”[tiab] OR “Hypothalamic pituitary axis”[tiab] OR “Hypothalamic pituitary adrenocortical system”[tiab] OR “pituitary adrenal cortical axis”[tiab] OR “Hypothalamo-pituitary-adrenal axis”[tiab] OR “hypothalamus–pituitary–adrenal axis”[tiab] OR “hypothalamo-pituitary-adrenal”[tiab] OR “major depression”[ti] OR “depressive illness”[ti] OR “major depressive disorder”[ti] OR “depressed”[ti] OR “antidepressant treatment”[ti] OR “salivary cortisol”[ti] OR “reliability”[ti] OR “cortisol response”[ti]) AND (“humans”[MeSH Terms] AND English[lang] AND “adult”[MeSH Terms] AND (“1982”[PDAT] : “2010/06/30”[PDAT])) NOT review[ptyp].

## References

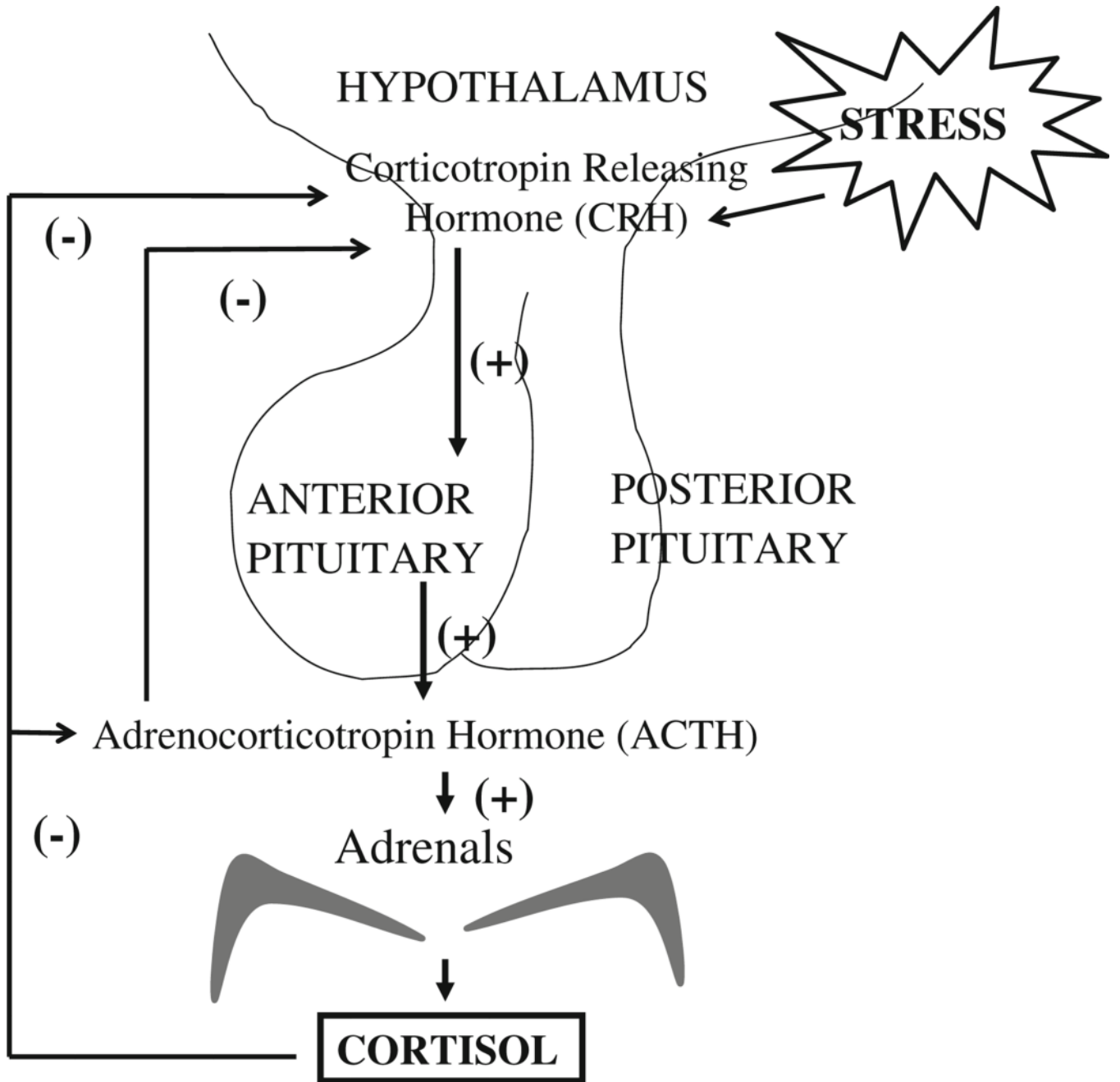
1. McEwen BS. Physiology and neurobiology of stress and adaptation: central role of the brain. *Physiol Rev.* 2007; 87:873–904. [PubMed: 17615391]
2. Juster RP, McEwen BS, Lupien SJ. Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neurosci Biobehav Rev.* 2010; 35:2–16. [PubMed: 19822172]
3. Rugulies R. Depression as a predictor for coronary heart disease A review and meta-analysis. *Am J Prev Med.* 2002; 23:51–61. [PubMed: 12093424]
4. Mezuk B, Eaton WW, Albrecht S, Golden SH. Depression and type 2 diabetes over the lifespan: a meta-analysis. *Diabetes Care.* 2008; 31:2383–2390. [PubMed: 19033418]
5. Golden SH, Lazo M, Carnethon M, Bertoni AG, Schreiner PJ, Roux AV, Lee HB, Lyketsos C. Examining a bidirectional association between depressive symptoms and diabetes. *JAMA.* 2008; 299:2751–2759. [PubMed: 18560002]
6. Golden SH. A review of the evidence for a neuroendocrine link between stress, depression and diabetes mellitus. *Curr Diabetes Rev.* 2007; 3:252–259. [PubMed: 18220683]
7. Szklo, M.; Nieto, FJ. *Epidemiology beyond the Basics.* Sudbury: Jones and Bartlett Publishers; 2004. Quality assurance and control; p. 343-404.
8. Derr RL, Cameron SJ, Golden SH. Pre-analytic considerations for the proper assessment of hormones of the hypothalamic-pituitary axis in epidemiological research. *Eur J Epidemiol.* 2006; 21:217–226. [PubMed: 16547837]
9. Hellhammer DH, Wust S, Kudielka BM. Salivary cortisol as a biomarker in stress research. *Psychoneuroendocrinology.* 2009; 34:163–171. [PubMed: 19095358]
10. Raff H. Utility of salivary cortisol measurements in Cushing’s syndrome and adrenal insufficiency. *J Clin Endocrinol Metab.* 2009; 94:3647–3655. [PubMed: 19602555]
11. Groschl M, Rauh M. Influence of commercial collection devices for saliva on the reliability of salivary steroids analysis. *Steroids.* 2006; 71:1097–1100. [PubMed: 17070563]
12. Garde AH, Hansen AM. Long-term stability of salivary cortisol. *Scand J Clin Lab Invest.* 2005; 65:433–436. [PubMed: 16081365]
13. Levine A, Zagoory-Sharon O, Feldman R, Lewis JG, Weller A. Measuring cortisol in human psychobiological studies. *Physiol Behav.* 2007; 90:43–53. [PubMed: 17055006]
14. Broderick JE, Arnold D, Kudielka BM, Kirschbaum C. Salivary cortisol sampling compliance: comparison of patients and healthy volunteers. *Psychoneuroendocrinology.* 2004; 29:636–650. [PubMed: 15041086]
15. Badrick E, Kirschbaum C, Kumari M. The relationship between smoking status and cortisol secretion. *J Clin Endocrinol Metab.* 2007; 92:819–824. [PubMed: 17179195]

16. Granger DA, Hibel LC, Fortunato CK, Kapelewski CH. Medication effects on salivary cortisol: tactics and strategy to minimize impact in behavioral and developmental science. *Psychoneuroendocrinology*. 2009; 34:1437–1448. [PubMed: 19632788]
17. Coste J, Strauch G, Letrait M, Bertagna X. Reliability of hormonal levels for assessing the hypothalamic-pituitary-adrenocortical system in clinical pharmacology. *Br J Clin Pharmacol*. 1994; 38:474–479. [PubMed: 7893592]
18. Harris TO, Borsanyi S, Messari S, Stanford K, Cleary SE, Shiers HM, Brown GW, Herbert J. Morning cortisol as a risk factor for subsequent major depressive disorder in adult women. *Br J Psychiatr*. 2000; 177:505–510.
19. Viardot A, Huber P, Puder JJ, Zulewski H, Keller U, Muller B. Reproducibility of nighttime salivary cortisol and its use in the diagnosis of hypercortisolism compared with urinary free cortisol and overnight dexamethasone suppression test. *J Clin Endocrinol Metab*. 2005; 90:5730–5736. [PubMed: 16014408]
20. Masserini B, Morelli V, Bergamaschi S, Ermetici F, Eller-Vainicher C, Barbieri AM, Maffini MA, Scillitani A, Ambrosi B, Beck-Peccoz P, Chiodini I. The limited role of midnight salivary cortisol levels in the diagnosis of subclinical hypercortisolism in patients with adrenal incidentaloma. *Eur J Endocrinol*. 2009; 160:87–92. [PubMed: 18835977]
21. Pruessner JC, Wolf OT, Hellhammer DH, Buske-Kirschbaum A, von Auer K, Jobst S, Kaspers F, Kirschbaum C. Free cortisol levels after awakening: a reliable biological marker for the assessment of adrenocortical activity. *Life Sci*. 1997; 61:2539–2549. [PubMed: 9416776]
22. Edwards S, Clow A, Evans P, Hucklebridge F. Exploration of the awakening cortisol response in relation to diurnal cortisol secretory activity. *Life Sci*. 2001; 68:2093–2103. [PubMed: 11324714]
23. Kraemer HC, Giese-Davis J, Yutsis M, O'Hara R, Neri E, Gallagher-Thompson D, Taylor CB, Spiegel D. Design decisions to optimize reliability of daytime cortisol slopes in an older population. *Am J Geriatr Psychiatr*. 2006; 14:325–333.
24. Ranjit N, Diez-Roux AV, Sanchez B, Seeman T, Shea S, Shrager S, Watson K. Association of salivary cortisol circadian pattern with cynical hostility: multi-ethnic study of atherosclerosis. *Psychosom Med*. 2009; 71:748–755. [PubMed: 19592518]
25. Hruschka DJ, Kohrt BA, Worthman CM. Estimating between-and within-individual variation in cortisol levels using multilevel models. *Psychoneuroendocrinology*. 2005; 30:698–714. [PubMed: 15854786]
26. Knutsson U, Dahlgren J, Marcus C, Rosberg S, Bronnegard M, Stierna P, Albertsson-Wikland K. Circadian cortisol rhythms in healthy boys and girls: relationship with age, growth, body composition, and pubertal development. *J Clin Endocrinol Metab*. 1997; 82:536–540. [PubMed: 9024250]
27. Schmidt-Reinwald A, Pruessner JC, Hellhammer DH, Federenko I, Rohleder N, Schurmeyer TH, Kirschbaum C. The cortisol response to awakening in relation to different challenge tests and a 12-hour cortisol rhythm. *Life Sci*. 1999; 64:1653–1660. [PubMed: 10328525]
28. Arnaldi G, Angeli A, Atkinson AB, Bertagna X, Cavagnini F, Chrousos GP, Fava GA, Findling JW, Gaillard RC, Grossman AB, Kola B, Lacroix A, Mancini T, Mantero F, Newell-Price J, Nieman LK, Sonino N, Vance ML, Giustina A, Boscaro M. Diagnosis and complications of Cushing's syndrome: a consensus statement. *J Clin Endocrinol Metab*. 2003; 88:5593–5602. [PubMed: 14671138]
29. Ljung T, Andersson B, Bengtsson BA, Bjorntorp P, Marin P. Inhibition of cortisol secretion by dexamethasone in relation to body fat distribution: a dose-response study. *Obes Res*. 1996; 4:277–282. [PubMed: 8732962]
30. Huizenga NA, Koper JW, de Lange P, Pols HA, Stolk RP, Grobbee DE, de Jong FH, Lamberts SW. Interperson variability but intraperson stability of baseline plasma cortisol concentrations, and its relation to feedback sensitivity of the hypothalamo-pituitary-adrenal axis to a low dose of dexamethasone in elderly individuals. *J Clin Endocrinol Metab*. 1998; 83:47–54. [PubMed: 9435415]
31. Golden SH, Malhotra S, Wand GS, Brancati FL, Ford D, Horton K. Adrenal gland volume and dexamethasone-suppressed cortisol correlate with total daily salivary cortisol in African-American women. *J Clin Endocrinol Metab*. 2007; 92:1358–1363. [PubMed: 17284636]

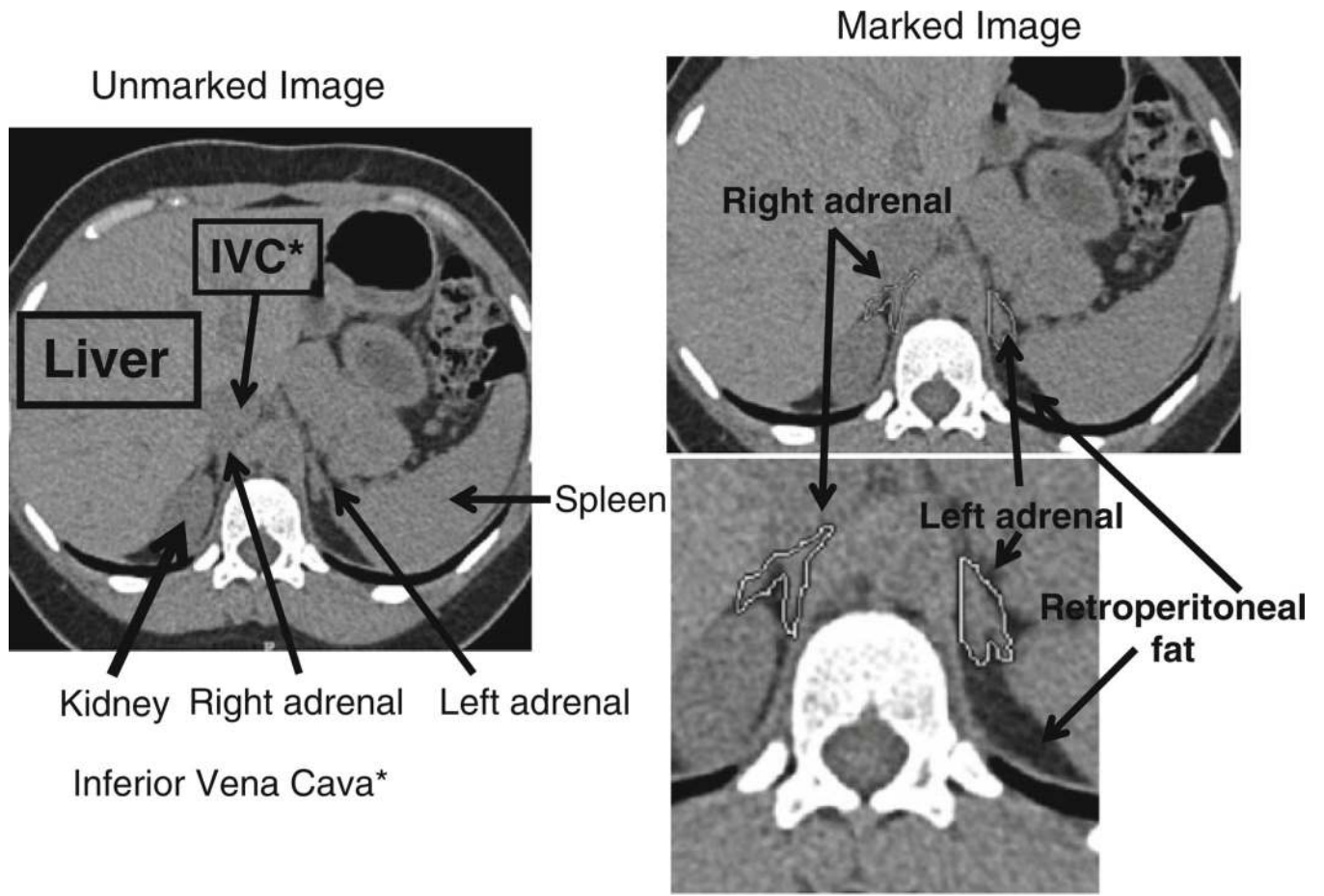
32. Yanovski SZ, Yanovski JA, Gwirtsman HE, Bernat A, Gold PW, Chrousos GP. Normal dexamethasone suppression in obese binge and nonbinge eaters with rapid weight loss. *J Clin Endocrinol Metab.* 1993; 76:675–679. [PubMed: 8445025]
33. Wingenfeld K, Lange W, Wulff H, Bera C, Beblo T, Saavedra AS, Mensebach C, Driessen M. Stability of the dexamethasone suppression test in borderline personality disorder with and without comorbid PTSD: a one-year follow-up study. *J Clin Psychol.* 2007; 63:843–850. [PubMed: 17674401]
34. Reynolds RM, Bendall HE, Whorwood CB, Wood PJ, Walker BR, Phillips DI. Reproducibility of the low dose dexamethasone suppression test: comparison between direct plasma and salivary cortisol assays. *Clin Endocrinol (Oxf).* 1998; 49:307–310. [PubMed: 9861320]
35. Gozansky WS, Lynn JS, Laudenslager ML, Kohrt WM. Salivary cortisol determined by enzyme immunoassay is preferable to serum total cortisol for assessment of dynamic hypothalamic-pituitary-adrenal axis activity. *Clin Endocrinol (Oxf).* 2005; 63:336–341. [PubMed: 16117823]
36. Harris B, Watkins S, Cook N, Walker RF, Read GF, Riad-Fahmy D. Comparisons of plasma and salivary cortisol determinations for the diagnostic efficacy of the dexamethasone suppression test. *Biol Psychiatr.* 1990; 27:897–904.
37. Imaki T, Naruse M, Takano K. Adrenocortical hyperplasia associated with ACTH-dependent Cushing's syndrome: comparison of the size of adrenal glands with clinical and endocrinological data. *Endocr J.* 2004; 51:89–95. [PubMed: 15004414]
38. Rubin RT, Phillips JJ, Sadow TF, McCracken JT. Adrenal gland volume in major depression. Increase during the depressive episode and decrease with successful treatment. *Arch Gen Psychiatr.* 1995; 52:213–218. [PubMed: 7872849]
39. Cooper MS, Stewart PM. Diagnosis and treatment of ACTH deficiency. *Rev Endocr Metab Disord.* 2005; 6:47–54. [PubMed: 15711914]
40. Klose M, Lange M, Kosteljanetz M, Poulsgaard L, Feldt-Rasmussen U. Adrenocortical insufficiency after pituitary surgery: an audit of the reliability of the conventional short synacthen test. *Clin Endocrinol (Oxf).* 2005; 63:499–505. [PubMed: 16268800]
41. Rubin RT, Phillips JJ, McCracken JT, Sadow TF. Adrenal gland volume in major depression: relationship to basal and stimulated pituitary-adrenal cortical axis function. *Biol Psychiatr.* 1996; 40:89–97.
42. Amsterdam JD, Marinelli DL, Arger P, Winokur A. Assessment of adrenal gland volume by computed tomography in depressed patients and healthy volunteers: a pilot study. *Psychiatr Res.* 1987; 21:189–197.
43. Brauckhoff M, Stock K, Stock S, Lorenz K, Sekulla C, Brauckhoff K, Thanh PN, Gimm O, Spielmann RP, Dralle H. Limitations of intraoperative adrenal remnant volume measurement in patients undergoing subtotal adrenalectomy. *World J Surg.* 2008; 32:863–872. [PubMed: 18224482]
44. Grant LA, Napolitano A, Miller S, Stephens K, McHugh SM, Dixon AK. A pilot study to assess the feasibility of measurement of adrenal gland volume by magnetic resonance imaging. *Acta Radiol.* 2010; 51:117–120. [PubMed: 20088645]
45. Anagnostis P, Athyros VG, Tziomalos K, Karagiannis A, Mikhailidis DP. Clinical review: the pathogenetic role of cortisol in the metabolic syndrome: a hypothesis. *J Clin Endocrinol Metab.* 2009; 94:2692–2701. [PubMed: 19470627]
46. Greden JF, Gardner R, King D, Grunhaus L, Carroll BJ, Kronfol Z. Dexamethasone suppression tests in antidepressant treatment of melancholia. The process of normalization and test-retest reproducibility. *Arch Gen Psychiatr.* 1983; 40:493–500. [PubMed: 6340634]
47. Holsboer F, Liebl R, Hofschuster E. Repeated dexamethasone suppression test during depressive illness. Normalisation of test result compared with clinical improvement. *J Affect Disord.* 1982; 4:93–101. [PubMed: 6213695]
48. Heuser I, Yassouridis A, Holsboer F. The combined dexamethasone/CRH test: a refined laboratory test for psychiatric disorders. *J Psychiatr Res.* 1994; 28:341–356. [PubMed: 7877114]
49. Kunzel HE, Binder EB, Nickel T, Ising M, Fuchs B, Majer M, Pfennig A, Ernst G, Kern N, Schmid DA, Uhr M, Holsboer F, Modell S. Pharmacological and nonpharmacological factors influencing hypothalamic-pituitary-adrenocortical axis reactivity in acutely depressed psychiatric

in-patients, measured by the Dex-CRH test. *Neuropsychopharmacology*. 2003; 28:2169–2178.  
[PubMed: 12931142]





**Fig. 1.** Physiology of the hypothalamic–pituitary–adrenal axis



**Fig. 2.**  
CT scan showing anatomic location of the left and right adrenal glands

Reliability, advantages and disadvantages of hypothalamic-pituitary-adrenal (HPA) axis measures for consideration of measuring chronic cortisol burden in population-based studies

Table 1

Measure	Reliability measure*	Study characteristics	Time between repeat measures	Advantages	Disadvantages
8 a.m. salivary cortisol (single measurement)	R = 0.18 (17)	20 Normal healthy male volunteers	1–5 weeks	Easy to perform in population-based and intervention studies in free-living state without assistance from medical personnel	Poor between visit reliability
	$r_s = 0.41$ (18)	Age range 21–29 years			Likely does not reflect chronic endogenous cortisol exposure and HPA axis tone
	R = 0.47 (19)	116 Women vulnerable to depression but not currently depressed Mean age 38.5 years	6 months	Assesses free and not bound cortisol	
		20 Normal healthy volunteers (10 men, 10 women) Mean age (range): 32.5 (20–58) years Mean BMI 22.1 kg/m <sup>2</sup>	3 days of sequential sampling (although exact time between sample collections not specified)	Stable for several days prior to processing	
	R = 0.27	48 Men and women in Atherosclerosis Risk In Communities Carotid MRI Study Mean age 71.3 ± 5.4 years 40% Women; 73% White, 27% Black	81.7 days		
11 p.m. salivary cortisol (single measurement)	R = 0.78 (19)	20 Normal healthy volunteers (10 men, 10 women) Mean age (range): 32.5 (20–58) years Mean BMI 22.1 kg/m <sup>2</sup>	3 days of sequential sampling (although exact time between sample collections not specified)	Same advantages of 8 a.m. salivary cortisol but also greater between visit reliability over several days in healthy individuals	Lower sensitivity in detecting subclinical hypercortisolism than dexamethasone suppressed cortisol or UFC
Multiple daily salivary cortisol measurements from diurnal curve	$r = 0.647$ (22)	42 Healthy volunteers (32 women—27 non-smokers, 5 smokers; 10 males—9 non-smokers, 1 smoker) Mean age (range) 35 (23–59) years	1 day	May reflect integrated cortisol secretion through correlation with other measures of chronic HPA axis tone (e.g. DST, adrenal gland volume)	Requires collection of multiple samples throughout the day at several time points, which may reduce compliance with field sampling
	$r_s = 0.63$ (23) Predicted for 2 days of sample collection: $r_s = 0.78$	Older adults mean age 61.5 ± 11.7 years; 88% female; 12% Hispanic white and 78% non-Hispanic white	1 day	Allows collection of multiple samples in a field setting	
	Predicted for 3 days of sample collection: $r_s = 0.84$			Successfully performed in a large, population-based study	
	$r = 0.65$ – $0.72$ (24)	936 Men and women from the Multi-Ethnic Study of Atherosclerosis Mean age 61.4 years; mean BMI 29 kg/	3 days of sequential sampling with one evening sample	May be able to characterize the diurnal curve with awake and late evening sample	

Measure	Reliability measure*	Study characteristics	Time between repeat measures	Advantages	Disadvantages
		m <sup>2</sup> 48% Women; 52% men	day of separation between measurements	Data from multiple time points of collection can be used to in multilevel, mixed, or hierarchical linear models to account for within and between individual differences in cortisol curves	
Awakening cortisol	r = 0.39–0.67 (21)	19.2% White; 28.3% Black; 52.5% Hispanic Children (21 boys; 21 girls; mean age 11.2 ± 2 years)	3 days of sequential sampling with one day of separation between measurements	Higher intra-individual stability than single morning cortisol measurement Can uncover subtle changes in HPA tone due to chronic stress Reflects adrenal capacity to respond to stress	Affected by other factors such as sex, smoking, and oral contraceptives Participants require instruction about collection
	r = 0.42–0.65 (21)	Young adults (35 men; 35 women; mean age 26.5 ± 6.3 years)	3 sample collections separated by 1 week		Accurate timing of sample collection required
	r = 0.58 (21)	Elderly adults (20 men; 20 women; mean age 70.4 ± 5.7 years)	1 day		
	r = 0.336–0.52 (22)	Young adults (10 men; 32 women; mean age 35 years)	1 day		
Dexamethasone-suppressed cortisol	r = 0.42 (8 a.m. morning cortisol following Dexamethasone suppression) (33)	13 Women with Borderline Personality Disorder (8 also had Post-Traumatic Stress Disorder)	1 year	Well-established tool in clinical endocrinology for assessing HPA axis tone Reflects chronic as opposed to acute HPA axis tone and cortisol exposure	Requires dexamethasone administration Requires timed administration of dexamethasone and timed collection of cortisol sample
		Mean age 33.5 years		Can be performed with collection of salivary cortisol in large populations	Use with post-dexamethasone salivary cortisol may not be as reliable in small populations
	r = 0.56 (evening [4 p.m.] cortisol following dexamethasone suppression) (33)	Used 0.5 mg dex suppression test			

Measure	Reliability measure*	Study characteristics	Time between repeat measures	Advantages	Disadvantages
Adrenal gland volume	Between visit repeatability: $r_s = 0.66$ (30) Non-contrast adrenal CT scan Inter-reader reliability: Left adrenal gland: $r_s = 0.67-0.71$ Right adrenal gland: $r_s = 0.47-0.70$	164 Healthy elderly from the Rotterdam study Mean age $67.7 \pm 5.6$ years in 102 men Mean age $65.8 \pm 6.1$ years in 114 women Used 1 mg dex suppression test at baseline and 0.25 mg dex suppression test at follow-up 28 Healthy African-American women Mean age $33.4 \pm 7.9$ years	2-5 years  1-2 years (between repeat readings of scans)	Non-invasive measure that reflects chronic endogenous cortisol exposure and HPA axis tone  Can be performed during CT scans or MRI that study participants might be already receiving	Radiation exposure for CT scan (~320 mrem) Expensive Volume calculations are labor intensive  Inter-reader reliability may depend on the technical skill of the reader Variation between scanners may introduce additional decrease in reliability
	Adrenal CT scan with intravenous contrast Between visit intra-individual reliability of adrenal volume remnant: $r = 0.84$ (43)	Pheochromocytoma patients following subtotal adrenalectomy (n = 22; mean age $42.8 \pm 16.3$ years; 55% women)	3 months (between repeat scans)		Use of contrast with CT scan (i.e. renal complications, allergic reactions) Visualization of the adrenals affected by degree of retroperitoneal fat present on non-contrast CT or MRI
	Magnetic resonance imaging (MRI) (44) Between visit variation 0.14 cm <sup>3</sup> (5% of 3 cm <sup>3</sup> adrenal gland) Between investigator variation 0.28 cm <sup>3</sup> (9% of 3 cm adrenal gland)	4 Healthy individuals (3 men; 1 woman; age range 18-45 years; body-mass index 18-25 kg/m <sup>2</sup> )	7 days (between repeat scans)	MRI: No radiation exposure	

R. Intraclass correlation coefficient; r, Pearson's linear correlation coefficient;  $r_s$ , Spearman's ordinal correlation coefficient

\* Summary statistic represents between visit intra-individual reliability for all HPA axis measures except for reliability of adrenal gland volume by non-contrast CT scan, where the summary statistic represents inter-reader reliability

**Table 2**

Reliability coefficients (R) for 8 a.m. salivary cortisol from 146 participants in the Atherosclerosis Risk In Communities Carotid MRI Study

	N	Reliability coefficient (R)	Coefficient of variation (%)
Within visit quality control replicates	98	0.86	19.2
Repeat visit quality control replicates	48	0.27	48.8

**Table 3**

Inter-reader reliability ( $r_s$ ) of left and right adrenal gland volume from 28 healthy African-American women by body-mass index (BMI) status

	N	Inter-reader reliability coefficient ( $r_s$ )	
		Left adrenal gland volume	Right adrenal gland volume
BMI $\geq 25$ kg/m <sup>2</sup>	17	0.71 ( $P=0.001$ )	0.47 ( $P=0.06$ )
BMI $< 25$ kg/m <sup>2</sup>	11	0.67 ( $P=0.02$ )	0.70 ( $P=0.02$ )