Salivary cortisol sampling compliance: comparison of patients and healthy volunteers

Joan E. Broderick a,*, Daniel Arnold a, Brigitte M. Kudielka b, Clemens Kirschbaum c

a Department of Psychiatry and Behavioral Science, Patnam Hall, Stony Brook University, Stony Brook, NY 11794-8790, USA
b Department of Behavioural Sciences, Swiss Federal Institute of Technology, Zurich, Switzerland
c Department of Experimental Psychology, University of Duesseldorf, Duesseldorf, Germany

Received 27 December 2002; received in revised form 5 April 2003; accepted 16 April 2003

Abstract

Objective: Problems of compliance with in vivo data collection and treatment protocols have been identified. This study investigated compliance with salivary cortisol sampling in a 7-day protocol. Impact of non-compliance on cortisol data was evaluated. Methods: Female fibromyalgia patients were matched with healthy female volunteers and randomized to Aware or Unaware conditions regarding objective monitoring of their sampling compliance. The protocol entailed collecting five saliva samples at prescribed times on each of 7 consecutive days. Participants self-reported time of each sample, and electronic monitor caps provided an objective date and time stamp of each sample. Results: Objective compliance among participants unaware of monitoring was 71%, though their self-reported compliance was 93%. Aware participants’ objective compliance was 90% which was consistent with self-reported compliance of 93%. Within-subject comparison of early morning rise and day slope of cortisol for compliant and non-compliant samples found significant differences with non-compliant samples resulting in flatter slopes. Patients were somewhat more compliant than healthy volunteers. Slight decrements in compliance were found for the afternoon sample (1600 h) and for the last 3 days of sampling. Compliance did not differ on weekdays versus weekends. Conclusions: Self-report of compliance in a salivary cortisol sampling protocol substantially overestimates actual compliance in the absence of objective monitoring. Non-compliance with the sampling protocol was lower in the Aware condition.

* Corresponding author. Tel.: +1-631-632-8083; fax: +1-631-632-3165.

Abbreviations:
FM, fibromyalgia

E-mail address: joan.broderick@stonybrook.edu (J.E. Broderick).
protocol results in cortisol data that significantly differs from compliant data. Awareness of electronic monitoring of sampling results in satisfactory compliance. © 2003 Elsevier Ltd. All rights reserved.

**Keywords:** Compliance; Adherence; Cortisol; Salivary cortisol sampling; Fibromyalgia

### 1. Introduction

Across the past 20 years there has been a prominent movement toward collection of data in the natural environment. While in vivo measurement has distinct advantages for some research questions, it also has the disadvantage of losing the control afforded by the laboratory or clinic. To the degree that patients must actively initiate or participate in measurement outside of the laboratory, factors such as patient motivation or ability can affect the quality of the data. Lack of compliance with the measurement protocol is potentially a significant cause of missing data and of data of questionable validity. For example, the problem has been found to be substantial in investigations of medication dosing compliance where missed doses and inappropriate multiple dosing are common and probably account for many reported treatment failures (Burney et al., 1996; Vanhove et al., 1996). Moreover, the evidence clearly suggests that patients attempt to appear compliant with their medication by generating self-reports of their behavior that inflate actual compliance (Rand et al., 1992; Spector et al., 1986). Similar problems have been observed in studies in which patients report glucose levels (Mazze et al., 1984) and blood pressure readings (Nordmann et al., 2000). Such dissembling places the health provider at a disadvantage in making treatment decisions and provides a researcher with data that are collected at a time deviating from the sampling protocol, thus potentially introducing significant error variance. In the absence of objective information on protocol compliance, unexpected treatment outcomes or research results are difficult to interpret with confidence.

Psychoendocrinology studies often involve collection of data in the field. With the advent of salivary assays, these studies have increased due to their non-invasive nature and the ease of sample collection by research participants (Kirschbaum et al., 1990). Typically, participants are given instruction in the procedures for collection of saliva, are provided with salivettes, and are directed to take samples over one or more days according to a sampling protocol. Participants self-report the date and time each sample was taken. This self-reported information is accepted as veridical by researchers. However, some recent evidence suggests that, as in medication studies, patients’ reports of their compliance with protocol may not be entirely accurate. Kudielka found in a 1-day protocol that 26% of participants collected at least one of six saliva samples outside of the designated sampling times, and the non-compliance significantly affected the cortisol profiles obtained (Kudielka et al., 2003). Sampling of hormones, such as cortisol, that vary quickly over time require accuracy of sampling time to yield valid conclusions (Wehr et al., 2001). Especially in the morning, when cortisol rises rapidly after awakening and then quickly falls, deviations in sam-
pling times can dramatically impact the slopes obtained (Kudielka et al., 2003; Pruessner et al., 1997).

It is reasonable to assume that participant characteristics may impact on level of compliance with a protocol. We hypothesized that patients with an illness would be more motivated in a research protocol to provide valid data in order to facilitate research on their disease than would healthy control individuals who are volunteering for less personal, more altruistic reasons. Thus, we designed this study to compare compliance between a sample of fibromyalgia (FM) patients and a matched, healthy control sample. FM is a chronic pain disorder of unknown etiology (Wolfe et al., 1990) and limited treatment options (Rossy et al., 1999), thus patients express a high degree of enthusiasm for research on the condition. We hypothesized that compliance would be greater in our FM patients than in our controls.

We also assumed that participants who know that they are being monitored for sampling compliance would provide more compliant data than participants who were unaware that they were being monitored. Thus, we randomly assigned participants into groups who were either aware or unaware of objective sampling compliance monitoring. The unaware participants would provide an estimate of the compliance behavior in a typical research protocol in which self-report of sampling times in an “honor system” is provided by participants.

Finally, we wanted to examine the pattern of sampling compliance for different times of day, on weekdays versus weekends, and as the sampling burden increased over the course of 7 days. Compliance might vary systematically on these time dimensions as has been found in treatment studies (Cramer et al., 1990).

1.1. Specific study questions

The following specific hypotheses will be explored. Hypothesis 1: self-reported times of sample collection will not be consistent with objective evidence of sample times. Hypothesis 2: participant awareness of objective monitoring of sampling times will produce greater compliance with the sampling protocol. Hypothesis 3: patients will provide more compliant samples than healthy control participants. Hypothesis 4: cortisol profiles will be significantly impacted by non-compliance (a) for the early morning rise and (b) for the morning to evening slope.

Secondary analyses will explore the following questions: (1) Will compliance degrade over the 7 days of sampling? (2) Will compliance vary by time-of-day of sampling? (3) Is compliance different on the weekends (Friday–Sunday) versus weekdays?

2. Method

2.1. Participants

Women with FM and healthy controls were recruited through newspaper announcements and fliers in an academic hospital. The stated purpose of the study
was to compare adrenal hormone levels of women with FM with matched, healthy women. Interested women were screened on the telephone for eligibility criteria including age (18–70), no significant problems with sight or hearing, fluent in English, not pregnant, no history of Schmidt’s or Cushing’s syndrome, no history of adrenal tumors or renal insufficiency, no current major depressive episodes or current bulimia, or no use of steroid medications, and physician-confirmed diagnosis of FM. Each control matched one FM patient on age ±5 years, menopausal status, smoker versus non-smoker, and use of oral contraceptives since these factors can influence cortisol levels. To be classified as a healthy control, a woman could not have any serious past or present medical conditions. Of the 58 FM women who were screened, 16 (28%) were ineligible, 6 (10%) were uninterested in the study, and 3 (5%) who were eligible could not be scheduled. Of the 45 control women who were screened, 8 (18%) were ineligible because they did not match a FM patient and another 2 (4%) were ineligible for other reasons, and 2 (4%) were uninterested in the study. The final sample included 33 FM patients and 33 controls.

2.2. Materials and procedure

FM patients were randomly assigned to either the Aware or Unaware condition. Each control was assigned to the same condition as the FM patient to whom she was matched. The entire Unaware condition was run prior to the Aware condition to prevent any contamination of knowledge of the purpose of the electronic monitoring cap by the unaware participants.

Study participants came into the research office as they were recruited, gave informed consent, completed the SF36v2, (Ware, 1993; Ware et al., 2000) and received instructions in the procedure for collecting saliva samples. They were provided with a booklet to take home that reviewed the timing and procedures for sampling and an index card to record daily the time they awoke and went to sleep. Each participant received a 40-dram vial packed by the experimenter with 38 cotton swabs (three extras included) and closed with an eDEM™ electronic monitor cap (Aardex Ltd., Switzerland). They were also given 38 labeled salivettes (Sarstedt, Newton, North Carolina) to store each used cotton swab and to write the time the sample was taken.

Participants were instructed to take five saliva samples at designated times for the next 7 consecutive days and to store them in the freezer or refrigerator. Starting day was not systematically controlled and varied across participants. The sampling protocol was: immediately upon awakening, 45 min after awakening, 1600, 1900, and 2200 h (or just before sleep if that came earlier). Participants were reminded to wash their hands before taking a sample and to rinse their mouth with water to remove food particles if they had just eaten. Instructions were given to refrain from eating or drinking anything for at least 30 min after awakening, to refrain from brushing teeth for 30 min prior to a sample, or drinking an acidic product (e.g., orange juice) 15 min prior to a sample.

The Unaware participants were told that the eDEM™ cap was designed to keep the cotton swabs sanitary and reduce humidity, and that it was vital to only open the cap to remove a cotton swab to take a sample. They were not aware of the date
and time stamp function of the cap. The Aware participants were told that the eDEM™ cap would record each time the cap was opened and closed to provide information about adherence to the sampling protocol.

Participants returned to the research office at the end of the week with their samples. A debriefing interview was conducted to assess the participants’ perceptions of their compliance with the sampling protocol and to allow for discussion of any other relevant events during the week of sampling including suspicion about the purpose of the eDEM™ cap.

Prior to each use of an eDEM™ cap by a participant, it was tested in a standardized manner to verify accurate recording of the occurrence and timing of opening and closing. No errors were observed, and the cap was always within 60 s of the reference computer clock.

2.3. Cortisol analysis

Salivary cortisol was assayed with a time-resolved immunoassay with fluorometric detection (DELFIA) as described by Dressendorfer et al. (1992). Intra- and interassay coefficients of variability were below 12%. The lower detection limit of this assay is 0.43 pg/well (i.e., cortisol levels of 1 nmol/l are reliably measured as being different from zero). Samples with coefficients of variance over 12% were rerun in a second assay (standard practice in clinical lab medicine). Extreme cortisol values (≥4 SD above the mean) for each sample time were eliminated (4 Awake samples, 4 Awake + 45 samples), and to obtain normalization of the distribution cortisol values were transformed (log 10) prior to analysis.

2.4. Statistical analysis

Data from the eDEM™ caps were uploaded via Powerview (Aardex Ltd., version 1.4.0; 2001) software into an Excel spreadsheet. The data were cleaned by visual inspection; “stray” opening events (i.e., not coinciding with a targeted assessment time) were removed, and multiple openings/closings around an assessment time were removed. In all cases, the opening that was most compliant was retained for analysis.

Two types of missing data were defined: (1) no saliva sample collected, and (2) no self-reported time of sample written on the salivette. A missing saliva sample is a breach of protocol, but it is excluded from all of the compliance analyses, since it is not the type of non-compliance that can be hidden from a researcher or clinician. The absence of a self-reported sample time precludes the ability to determine self-reported compliance for that sample, therefore it had to be set to missing. In the case of the Awake and Awake + 45 samples, the sample would be set to missing if there was no awake time recorded on the diary card.

For each sample, two types of compliance were determined: (1) self-reported sampling (Self-Reported Compliance), and (2) sampling according to the eDEM™ cap (Verified Compliance). Although there was no definitive way to verify compliance for the Awake and Awake + 45 samples in this study (i.e., there was no objective way to determine when the participants actually woke up), it was possible to compare
the time awake reported on the Sleep–Awake Diary Card with the time the Awake sample was taken. Nine participants did not provide these diary cards. Due to the rapid change in cortisol in the first few hours after awakening, a ±15-min compliance window was established for the Awake and Awake + 45 samples. For example, the Awake + 45 sample was compliant if it was taken between 30 and ±60 min after the awake time recorded on the diary card. A more liberal compliance window of 60 min was used for the last three samples, 1600, 1900, 2200 h (e.g., 1500–1700 h for the 1600 h sample), since cortisol changes much more slowly in the afternoon and evening than in the early morning. In the few instances when the participant retired to bed prior to 2200 h, the sample was excluded if the patient went to bed before 2100 h.

Self-reported compliance was determined by the time recorded on the salivette by the participant. It was calculated across 7 days for each participant as the number of compliant samples per self-report by participant (numerator) divided by the total number of samples with non-missing sampling times. If there were no missing data, the denominator would be 35 (five samples across 7 days).

Verified compliance was determined by the time recorded by the eDEM™ opening; if the cap was not opened for a sample, the sample was defined as non-compliant. It was calculated across 7 days for each participant as the number of compliant samples per eDEM™ cap (numerator) divided by the total number of non-missing sampling times.

Group differences on relevant demographic variables were controlled in the analyses. An alpha level for inclusion of a control variable was set liberally at 0.20.

For statistical testing of differences among compliance rates, the arcsine transformation was applied in order to reduce the floor/ceiling effects and to reduce skewness in the distribution of rates. SAS version 8.0 was used for data analysis.

3. Results

3.1. Participant characteristics

All 66 participants completed the protocol, however one participant’s data (FM-Aware condition) were not analyzed because she discarded her eDEM™ cap. Table 1 displays the demographic comparisons for FM patients versus Healthy Controls and for Aware versus Unaware groups. As expected, there was a difference in employment ($\chi^2(2,N = 65) = 3.81, p = 0.15$), disability status (Fisher’s exact test ($N = 65$), $p < 0.01$), and physical ($t(63) = 11.33, p < 0.001$) and mental ($t(63) = 4.60, p < 0.001$) health status on the SF36 between the FM patients and the Healthy Controls. Unexpectedly, the Controls had more education ($\chi^2(3,N = 65) = 9.96, p = 0.02$) and were more ethnically diverse (Fisher’s exact test ($N = 65$), $p = 0.05$) than the FM patients. Comparisons of Aware and Unaware participants revealed differences in age ($t(63) = 2.03, p < 0.05$) and education ($\chi^2(3,N = 65) = 6.58, p = 0.09$). Consequently, education, race, and age are used as covariates in the subsequent analyses.
## Table 1
Demographic characteristics of participants

<table>
<thead>
<tr>
<th>Health status</th>
<th>FM (N = 32)</th>
<th>Healthy controls (N = 33)</th>
<th>p level</th>
<th>Awareness of monitoring</th>
<th>Unaware (N = 34)</th>
<th>Aware (N = 31)</th>
<th>p level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (M, SD)</td>
<td>50.6 (9.8)</td>
<td>50.7 (10.6)</td>
<td>ns</td>
<td></td>
<td>48.3 (10.1)</td>
<td>53.3 (9.6)</td>
<td>0.05</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married/living with partner</td>
<td>66%</td>
<td>67%</td>
<td>ns</td>
<td></td>
<td>65%</td>
<td>68%</td>
<td>ns</td>
</tr>
<tr>
<td>Separated/divorced/never married/widowed</td>
<td>34%</td>
<td>33%</td>
<td></td>
<td></td>
<td>35%</td>
<td>32%</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>50%</td>
<td>15%</td>
<td>0.02</td>
<td></td>
<td>21%</td>
<td>45%</td>
<td>0.09</td>
</tr>
<tr>
<td>Some college</td>
<td>22%</td>
<td>27%</td>
<td></td>
<td></td>
<td>24%</td>
<td>26%</td>
<td></td>
</tr>
<tr>
<td>College graduate</td>
<td>16%</td>
<td>24%</td>
<td></td>
<td></td>
<td>29%</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Post-graduate</td>
<td>13%</td>
<td>33%</td>
<td></td>
<td></td>
<td>26%</td>
<td>19%</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>100%</td>
<td>85%</td>
<td>0.05</td>
<td></td>
<td>91%</td>
<td>94%</td>
<td>ns</td>
</tr>
<tr>
<td>Employment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part-time</td>
<td>31%</td>
<td>24%</td>
<td>0.15</td>
<td></td>
<td>26%</td>
<td>29%</td>
<td>ns</td>
</tr>
<tr>
<td>Full-time</td>
<td>34%</td>
<td>58%</td>
<td></td>
<td></td>
<td>47%</td>
<td>45%</td>
<td></td>
</tr>
<tr>
<td>Not employed</td>
<td>34%</td>
<td>18%</td>
<td></td>
<td></td>
<td>26%</td>
<td>26%</td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than $ 35 000</td>
<td>26%</td>
<td>13%</td>
<td>ns</td>
<td></td>
<td>18%</td>
<td>20%</td>
<td>ns</td>
</tr>
<tr>
<td>$ 35 000–$75 000</td>
<td>42%</td>
<td>50%</td>
<td></td>
<td></td>
<td>42%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>Greater than $ 75 000</td>
<td>32%</td>
<td>38%</td>
<td></td>
<td></td>
<td>39%</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>Disabled</td>
<td>25%</td>
<td>0%</td>
<td>0.01</td>
<td></td>
<td>15%</td>
<td>10%</td>
<td>ns</td>
</tr>
<tr>
<td>SF-36 physical health</td>
<td>34.9 (9.8)</td>
<td>54.6 (6.5)</td>
<td>0.001</td>
<td></td>
<td>43.4 (12.7)</td>
<td>46.3 (11.5)</td>
<td>ns</td>
</tr>
<tr>
<td>SF-36 mental health</td>
<td>42.1 (11.3)</td>
<td>52.9 (7.2)</td>
<td>0.001</td>
<td></td>
<td>47.5 (11.0)</td>
<td>47.7 (10.9)</td>
<td>ns</td>
</tr>
</tbody>
</table>
3.2. Missing samples

In order to determine if Health Status or Aware Status influenced the completeness of the data provided by the participants, the frequencies of missing data were examined. Missing saliva samples, missing self-report of time of sampling, and non-openings of cap were tabulated. A two (Health Status) by two (Awareness Status) ANCOVA with three covariates was computed for each type of missing data. There were no differences among the groups for missing cortisol samples or missing self-report of sampling times. However for non-openings of the cap at a sampling time, a main effect of Aware Status ($F(1,59) = 14.46, p < 0.003$) indicated that participants who were Unaware of monitoring had more instances of not opening the cap than those Aware of monitoring (adjusted means: 7.5 versus 1.3). A trend was observed for FM patients to have fewer non-openings of the cap than Controls ($F(1,59) = 3.48, p = 0.07; \text{adjusted means: 2.8 versus 6.0}$). The interaction term was not significant.

3.3. Compliance with saliva sampling: self-report and electronically verified

Mean compliance rates for each group were calculated as described above in the analysis section. Table 2 displays the rates of Self-Reported and Verified compliance for each group. A two (Health Status) by two (Aware Status) ANCOVA with three covariates was computed for Self-Reported compliance. Neither main effects (Health Status or Aware Status) nor the interaction was significant. Across all participants Self-Reported compliance was 93%. The same analysis was computed for Verified compliance, and a main effect of Aware Status was found ($F(1,58) = 15.19, p < 0.001$). The adjusted means for Verified compliance for the Aware and Unaware groups were 90 and 71%. No difference was found for Health Status, but the interaction was significant ($F(1,58) = 4.62, p = 0.04$) with the Controls being more influenced by awareness of monitoring (Fig. 1).

An analysis was conducted to determine if Self-reported and Verified sample times differed. A two (Aware status) by two (Health status) repeated measure (Self-report

<table>
<thead>
<tr>
<th>Compliance</th>
<th>Unaware</th>
<th>Aware</th>
</tr>
</thead>
<tbody>
<tr>
<td>FM (N = 17)</td>
<td>Control (N = 17)</td>
<td>FM (N = 15)</td>
</tr>
<tr>
<td>Self-reported</td>
<td>94% ± 0.13</td>
<td>92% ± 0.09</td>
</tr>
<tr>
<td></td>
<td>60–100%</td>
<td>71–100%</td>
</tr>
<tr>
<td>Verified</td>
<td>80% ± 0.18</td>
<td>62% ± 0.28</td>
</tr>
<tr>
<td></td>
<td>33–100%</td>
<td>9–100%</td>
</tr>
</tbody>
</table>

* Values are mean ± SD and range across participants.

b $n = 14$: one participant was excluded from this analysis because she did not provide any self-reported sampling times—perhaps due to misunderstanding the protocol instructions.
versus Verified compliance) ANCOVA yielded a significant main effect for the repeated measure \((F(1,55) = 4.20, p = 0.05)\). Participants’ Self-reported compliance was 93%, whereas Verified compliance was 81%. An interaction of Aware status and compliance report was also observed \((F(1,55) = 14.95, p < 0.001)\), see Fig. 2. No other main effects or interactions were found.

3.4. Compliance rates by time-of-day

The Self-Reported and Verified compliance rates for each sampling time were compared across groups to determine if sampling at certain times of day were more vulnerable to non-compliance. A two (Aware status) by two (Health status) repeated
measures (time-of-day) ANCOVA was computed for Self-Reported compliance and for Verified compliance. There were no differences in Self-reported compliance among the sample times or by either grouping factor. The same analyses were computed for Verified compliance and sample time was significant ($F(4,196) = 2.45$, $p = 0.05$) as was the main effect of Aware status ($F(1,49) = 11.07$, $p = 0.002$). Post hoc comparisons revealed that the 1600 h sample had lower compliance than the Awake ($F(1,49) = 5.35$, $p = 0.03$), the Awake + 45 sample ($F(1,49) = 4.50$, $p = 0.04$), and the 1900 h sample ($F(1,49) = 7.84$, $p = 0.01$). The Verified compliance rates for the five sampling times across all participants starting with the Awake sample were: 88, 79, 76, 82, 78%; for the Aware participants they were: 94, 90, 90, 92, 85%.

3.5. Compliance rates by day-of-week

A further analysis was conducted to determine if compliance varied by weekday and weekend. Weekdays were defined as Monday through Thursday, and weekends were defined as Friday through Sunday. Friday was included in the weekend, since activities (e.g., going out with friends) in the evening of Friday are more consistent with weekend than weekday. This analysis was a repeated measures ANCOVA conducted across Health Status and Aware Status. There were no significant differences in compliance in this analysis.

3.6. Compliance rates compared at beginning and end of week

To determine if compliance degraded over the course of the 7 consecutive days of data collection, a two (Aware status) by two (Health status) repeated measure (days) ANCOVA was computed for Verified compliance. A trend was observed for the repeated measure ($F(1,58) = 1.79$, $p = 0.10$) showing a decrease in compliance from 91% after the first day of data collection down to the upper 70s for the last 3 days. The other factors did not interact with days.

3.7. Impact of compliance on cortisol slopes

In order to determine the impact of compliance on cortisol slopes, a within-subject comparison of the cortisol early morning rise (Awake+45–Awake) for compliant versus non-compliant days was computed. Inclusion of a participant in this analysis required at least 1 day each of compliant and non-compliant rises (i.e., a compliant rise required both samples to be compliant, a non-compliant rise involved either or both samples being non-compliant). Compliant rises were averaged and compared with the average of non-compliant rises within-subject. Thirty participants had data to compare compliant and non-compliant days (21 Unaware, 9 Aware). A repeated measures ANCOVA yielded a significant difference in early morning cortisol rise between compliant and non-compliant samples ($F(1,29) = 7.31$, $p = 0.01$) with the compliant samples showing the expected rise (1.15–1.27 logged cortisol values for
Awake and Awake + 45), and the non-compliant samples showing no rise (1.20–1.19).

The impact of compliance on the full day’s slope of cortisol was also examined. The Awake + 45 sample is expected to yield the highest cortisol reading for the day, and the 2200 h is expected to be the lowest reading. Consequently, the day’s slope was computed based on the difference between the Awake + 45 and 2200 h samples. Inclusion of a participant in this analysis required at least 1 day each of compliant and non-compliant slopes as defined above for the early morning rise. Forty participants had both compliant and non-compliant slopes (23 Unaware, 17 Aware) to compare. A repeated measures ANCOVA yielded a significant difference in full day slope between compliant and non-compliant samples ($F(1,39) = 5.87, p = 0.02$) with the compliant samples showing a larger slope (1.26–0.33 logged cortisol) than the non-compliant samples (1.24–0.42).

4. Discussion

This study was designed to investigate whether self-reported compliance in a typical in vivo 7-day salivary cortisol sampling protocol is accurate, and the degree to which non-compliance affects the cortisol data obtained. As expected, self-reported compliance was high across all participants and did not differ between patients and controls. However, objective measurement of compliance yielded more sobering results. Half of our participants were naïve, thereby representing the conditions of a typical protocol wherein participants and researchers rely upon self-report, whereas the other half of our participants were informed of the monitoring function of the eDEM™ electronic monitor cap. Those participants who knew they were being electronically monitored provided sampling that was highly compliant and consistent with their self-reported sampling times. Conversely, those participants who were unaware of monitoring were much less compliant: their self-reports were significantly inflated to appear more compliant than they actually were. This was particularly true for the healthy controls whose overall verified compliance was 62% versus their self-report of 92%. Even the FM patients had a 14% discrepancy between self-reported and verified compliance. For some participants unaware of monitoring the discrepancy was profound with verified compliance as low as 11% and self-reported compliance 60%. These results confirm the hypothesis that participants’ self-reported compliance with salivary cortisol sampling cannot be taken as valid in the absence of participants’ knowledge that independent, objective monitoring is in place. We also hypothesized that patients with an active disease would provide more compliant data than healthy volunteers due to increased motivation. Relative to healthy volunteers, our patients did show evidence of greater compliance, but it was still significantly discrepant with their self-report.

Evidence of non-compliant sampling in a study does not de facto result in the conclusion that cortisol findings from such a study would be invalid. To address this possibility, we examined how compliance impacted on cortisol’s early morning rise and slope for the day. These phenomena are commonly analyzed in studies probing
cortisol. The strength of our 7-day study design came from having sufficient repeated measures to permit analysis of the effect of compliance on the slopes obtained within-subjects. More than half of the sample had both compliant and non-compliant days for each of these measures. For both early morning rise and day slope, the change in cortisol was significantly greater for the compliant samples compared with the non-compliant samples. This was the case even with our liberal compliance windows of ±15 min for the morning samples and ±60 min for the afternoon and evening samples. These data are evidence that the level of non-compliance observed in this study is sufficient to impact substantively on observed cortisol patterns.

Other compliance analyses addressed the questions of time-of-day, day-of-week, and trend over days. The 1600 h sample was the most vulnerable to non-compliance, perhaps due to the rush of ending the workday, transporting children to activities, or commuting. Somewhat surprising, there was no evidence that compliance was worse on weekends when routines are relaxed, and people may be less time conscious. With regard to erosion of compliance over days of the protocol, we found a non-significant trend. Thus, we conclude that a 7-day protocol entailing five samples per day is feasible without significant compliance problems for specific sampling times or across days.

Non-compliance can come in the form of non-adherence to the sampling protocol, but with accurate self-report of the incorrect sampling time. Or, more perilously, it can come in the form of non-adherence with protocol and inaccurate, falsified self-report of sampling time. The inability of researchers to identify non-compliant samples with falsified sampling times results in increased error in their data and yields results that are significantly different from that collected within the sampling parameters. As shown in this and a previous study (Kudielka et al., 2003), resulting cortisol patterns are significantly altered by including samples with falsified sample times. Presumably, this can have important consequences for the ability to detect cortisol–behavior relationships. The error induced by falsified compliance may significantly diminish the ability to reveal existing relationships between cortisol and psychological or behavioral parameters. To the extent that falsified compliance is systematically associated with the phenomenon being investigated, true relationships could be obscured or incorrect relationships could be accepted. A hypothetical study investigating the relationship between stress and cortisol demonstrates this possibility. If higher stress was associated with lower cortisol sampling compliance and false reporting, then the higher stressed individuals would provide more invalid data than less stressed individuals. This might yield the following data: the higher stressed individuals’ ‘Awake’ cortisol might be determined to be significantly higher than lower stressed individuals because their samples were taken many minutes after awakening (as cortisol is rising toward the morning peak). Thus, this example study could erroneously conclude that high stress individuals have higher Awake cortisol and flatter morning rises than low stress individuals.

One might assert that error introduced by non-compliance could be overcome by increasing statistical power. Increasing the N or the number of measurements are typical strategies for dealing with suboptimal reliability of measurement. However, this strategy is based on the assumption that the error introduced by non-compliance
is random, but that is incorrect. Rather, it is a validity problem; the error is systemic, not random. Increasing the $N$ or the number of samples collected will yield the same level and direction of error. In fact, increasing statistical power would only increase the researcher’s confidence in a false result. Thus, the problem cannot be solved with that strategy.

An encouraging message to be taken from this study is that satisfactory compliance can be achieved when objective monitoring is employed. For all participants we observed verified compliance of about 90%, which we think is quite reasonable for a 1-week study in the natural environment. In previous work employing a 1-day protocol, we observed 97% compliance among participants aware of monitoring (Kudielka et al., 2003). Studies examining compliance with symptom diaries also find compliance in the low 90s for monitored, electronic diaries over a week or more of data collection (Stone et al., 2002).

An issue that only became apparent to us as we prepared for data analysis is a problem with the objective verification of compliance with the Awake and Awake + 45 min samples. Whereas other samples are anchored to a set time-of-day, e.g., 1900 h, these first two samples are triggered by awakening in the morning. The rapid change in cortisol at this time (early morning rise) is only measured accurately if samples are taken within minutes of awakening. Though this task seems straightforward, in fact, it is probably just as vulnerable to non-compliance as any other sample. Since awakening is a behavioral event that is self-reported, the observed discrepancy between self-report and verified sampling described in this paper should be extended to the Awake sample. In fact, comparison of participants’ written log of time awake and the verified time of Awake sampling resulted in numerous discrepancies with Awake sample being taken sometimes up to 50 min before the time recorded on the awake log. This clearly suggests that something as simple as participants’ record of when they awoke in the morning is often in error and cannot be confidently used to determine compliance with morning sampling. Objective determination of when a participant awakens is difficult. Some researchers have begun to explore the use of actigraphy to differentiate sleep from awakening and have found that, though imperfect, it is more accurate than self-report (Eissa et al., 2001; Jean-Louis et al., 1999).

Several inherent factors in this study may limit the generalizability of results. First, all of our participants were women who were primarily middle-aged. Compliance among other demographic groups could vary. Second, this is an observational study, thus patient motivation may have been less than it would be in a treatment study. However, compliance with treatment regimens across many patient populations has been shown to be only fair (Boudes, 1998; Urquhart, 1996). Likewise, participants were not paid, and it is plausible that payment could increase motivation and compliance.

This study replicates and extends our previous work (Kudielka et al., 2003) documenting research participants’ propensity to inflate self-reports of compliance in a salivary cortisol sampling protocol, and this misinformation cannot be identified by researchers without objective monitoring. Furthermore, this observation is consistent with data derived from symptom diary protocols. Whereas patients’ self-reported
diary records indicated 95% compliance with the sampling protocol, verified compliance was only 20% (Stone et al., 2002). A subsequent study attempting to maximize verified compliance with symptom diary recording via signaling with wrist-watches and extensive training of patients found similarly inadequate compliance despite participants’ self-reports of good compliance (85–91% self-reported compliance versus 29–39% verified compliance) (Broderick et al., in press). It is apparent that participants wish to appear compliant resulting in self-reports that are misleading to researchers. It appears that the task of in vivo data collection is a difficult one to accomplish without the explicit accountability inherent in some monitoring devices, such as electronic instrumentation.

Fortunately, we also have evidence that when participants are aware of monitoring of their compliance, they are able to comply with the protocol in a satisfactory manner. Moreover, objective monitoring permits the researcher to identify samples collected outside of the protocol parameters and to make informed choices to prevent such data from contaminating their results.

5. Acknowledgments

We would like to thank Arthur Stone and Steven Grossman for their assistance in the data analysis, and Jaclyn Gutlieber and Steven Choi for assistance with research materials preparation. This research was supported in part by the Applied Behavioral Medicine Research Institute, Stony Brook University, New York.

References


Ware, J.E., 1993. SF-36 Health Survey: Manual and Interpretive Guide. The Health Institute, New England Medical Center, Boston.

