Comprehensive Psychoneuroendocrinology 9 (2022) 100108

Contents lists available at ScienceDirect

Comprehensive Psychoneuroendocrinology

journal homepage: www.sciencedirect.com/journal/comprehensive-psychoneuroendocrinology

The Cortisol Assessment List (CoAL) A tool to systematically document and evaluate cortisol assessment in blood, urine and saliva

Sebastian Laufer [a,*], Sinha Engel [a], Sonia Lupien [b], Christine Knaevelsrud [a], Sarah Schumacher [a,c]

[a] Division of Clinical Psychological Intervention, Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany
[b] Centre for Studies on Human Stress, Institut Universitaire en Santé mentale de Montréal, Psychiatry Department, Université de Montréal, Montréal, Canada
[c] Clinical Psychology and Psychotherapy, Department of Psychology, Faculty of Health, Health and Medical University, Potsdam, Germany

ARTICLE INFO

Keywords:
Cortisol
Documentation standards
Quality control
Reliability
Replicability

ABSTRACT

Background: The reliable assessment of cortisol is a necessary requirement to produce replicable research. Several recommendations to increase cortisol assessment reliability exist. However, cortisol assessment methodology is still rather heterogeneous. For this reason, the Cortisol Assessment List (CoAL) was created.
The CoAL can be used to guide researchers during the planning phase and document which measures were taken to increase cortisol data reliability in original studies. Moreover, the CoAL can be used to evaluate data quality in meta research. The items representing strategies to obtain reliable cortisol data can be weighted to indicate which are absolutely necessary to consider and which could be applied less restrictively in order to balance data quality and feasibility. In this paper, the construction process of the CoAL is described.
Methods: Item synthesis of the CoAL included a literature search to extract empirically based suggestions regarding the reliable assessment of cortisol. Estimates for the item weighting system were obtained by inviting experts in the field to participate in an online survey (n = 25). Inter-rater reliability (IRR) of the CoAL, was determined by letting independent raters use the CoAL to evaluate a set of randomly selected original studies (k = 90).
Results: The CoAL was divided into four subscales related to the *reporting of sampling procedures,* the consideration of *state covariates, trait covariates* and *exclusion criteria*. Survey results indicated high agreement among experts for most items (89%) with approximately half of the items in the CoAL being classified as *necessary* (Cortisol Awakening Response (CAR): 52%; basal cortisol: 52%; reactive cortisol: 44%) in order to obtain reliable cortisol data. Inter-rater agreement was very high (Cohen's Kappa = .98 - 0.99), indicating sufficient psychometric quality of the CoAL.
Discussion: The CoAL is the first tool to systematically plan, document and evaluate cortisol assessment. The survey results indicate that the majority of respondents are aware of essential requirements to increase data reliability. However, results were heterogeneous for some items, highlighting the need to start a process of developing a broad scientific consensus regarding reliable cortisol assessment. The implementation of the CoAL could be a first step in this direction. In conclusion, the CoAL reflects empirical evidence and expert knowledge regarding cortisol assessment and can be used as a flexible tool to plan and document empirical studies or evaluate cortisol data quality in meta research.

1. Introduction

Psychoneuroendocrinology opens the opportunity to complement behavioral or questionnaire data. Information on the (re)activity of the hypothalamic-pituitary-adrenal axis (HPA axis) by cortisol assessment can be of great value as it adds an objective biological dimension to psychological or health research. On the one hand, experimental studies can benefit from this, because cortisol data may explain variance in test performance and provide a deeper understanding of physiological mechanisms involved in stress processing. On the other hand, the relative ease with which these measures can be incorporated into ecologically valid study designs allows researchers to capture HPA axis functioning in naturalistic settings outside of the laboratory.

These advantages come with the responsibility to assess cortisol

* Corresponding author. Freie Universität Berlin, Schwendener Strasse 27, 14195, Berlin, Germany.
E-mail address: sebastian.laufer@fu-berlin.de (S. Laufer).

https://doi.org/10.1016/j.cpnec.2021.100108
Received 15 October 2021; Received in revised form 22 December 2021; Accepted 22 December 2021
Available online 28 December 2021
2666-4976/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

correctly to maximize the reliability[1] of the data and report measurement protocols transparently, increasing replicability.[2] To this end, a variety of confounding influences have been identified and guidelines related to measurement protocols have been established [1–6]. Still, assessment protocols and the respective reporting standards vary significantly between studies and meta research (i.e. research summarizing the existing literature, including meta-analyses, systematic reviews or expert consensus guidelines) aiming to summarize results often faces the obstacle of high heterogeneity (e.g. Refs. [7,8]). A recent review on the assessment of the Trier Social Stress Test (TSST) has also identified assessment heterogeneity and insufficient transparency in reporting practices as a cause for concern regarding replication efforts even in this highly standardized stress task [9]. Hence, looking at cortisol studies of a certain time frame (e.g., the last 20 years), one can be sure to find a methodological conglomerate that can make it hard to compare results. Some of the cortisol data will certainly have been collected and analyzed according to the proposed guidelines, while other data collection regiments will be outdated or at least not according to current best practice. Of course, this issue can be circumvented by strict exclusion criteria during study selection. However, even high-quality cortisol assessment studies often differ in minor ways that nevertheless could affect the results significantly. Hence, finding a way to systematically account for these differences seems reasonable and worth pursuing.

Next to the difficulties in meta research, the degree of heterogeneity in study designs also complicates the planning phase of studies assessing cortisol. Many researchers studying cortisol (re)activity comply with the majority of recommended procedures to ensure data (e.g. Refs. [10,11]). However, practical constraints may impede them from following *all* suggestions made in the guidelines or accounting for all confounding influences. The extent to which studies follow or deviate from these recommendations has not been investigated systematically, yet. Nevertheless, it seems like some recommendations have become standard procedure, while others remain exceptions. For example, in psychological intervention studies, the collection of the cortisol awakening response (CAR) over at least two consecutive days [12], appears to be implemented in the majority of studies, since the publication of the CAR guidelines by Stalder and colleagues (2016; see Ref. [7] for an overview). Conversely, the recommendation to electronically monitor adherence to sampling times in naturalistic environments [5,13] appears to be implemented less frequently in these studies. The exact reasons for this apparent difference in adherence to two recommendations given in the same set of guidelines are hard to determine. It might be due to somewhat conflicting findings regarding the influence of noncompliance with correct sampling times on cortisol levels, with some studies suggesting a clear influence of noncompliance on diurnal cortisol profiles (e.g. Refs. [14,15]), while others do not [16]. Another reason could be financial constraints, since electronic monitoring devices constitute a further burden on the often rather restricted study budget. This clearly shows the need to transparently report, which recommendations with regard to sampling procedures are being followed and to investigate their relative importance in the field as indicated by the frequencies of their implementation.

Apart from decisions regarding sampling procedures, controlling for all of the potentially confounding influences may prove difficult in cortisol assessment. This issue is certainly not unique to psychoneuroendocrinology, but a general problem of psychological research [17–19]. However, it seems to be especially relevant in this line of research, since cortisol assessment in urine, blood and saliva not only has to account for general demographic differences like age, sex or body mass index (BMI), but also influences on the day of sampling like food, drink and caffeine intake or smoking [6]. Employing very strict sampling protocols (i.e., not eating, drinking, smoking or sleeping 60 min prior to sample collection) in order to account for the latter group of confounders can be problematic for various reasons. Firstly, complying with very restrictive sampling protocols may deter individuals from participating in studies or lead to high participant burden, increasing the probability of selection bias and/or drop-out (for a discussion, see Refs. [1,6,20]). Secondly, individuals' routines may be changed by the restrictions to such an extent that the assessment days do not reflect typical days in their lives anymore, which might introduce noise to the data and limit external validity, even though this has to our knowledge not been systematically investigated. Hence, finding the right balance between controlling for a sufficient number of influences to ensure good data quality and ensuring adherence to sampling protocols can be quite challenging, especially when dealing with vulnerable populations like individuals suffering from somatic or mental illnesses [8,21–23].

As a result of the issues raised above, many researchers studying cortisol have developed their own "in-house" data collection guidelines and there is at this point "… no gold standard for a sampling protocol …" as stated in a recently published paper on this issue ([24]; p. 3). The individual assessment guidelines used are often based on empirical evidence, overlap in many areas and it has been shown, that the intra-class correlation between laboratories analyzing cortisol is very high [25]. Nevertheless, different laboratories and researchers hardly ever use the exact same assessment protocols, complicating the comparison of results. Furthermore, the degree of detail to which researchers report these protocols in their publications varies significantly, exacerbating the replication problems [26,27].

1.1. Study rationale

To conquer this problem in meta research on cortisol, we had developed a tool (from here on referred to as the *Quality Estimation Checklist*) to gather information on the consideration of confounders within cortisol assessment and used it in two meta-analyses and a systematic review [7,8,28].

The aim of the current project was to ensure that the Quality Estimation Checklist meets the minimum requirements of psychometric quality control by establishing the inter-rater-reliability (IRR) across a broad variety of study designs. In connection to this, we reevaluated the existing tool and extended it where needed by adding more detailed items. Moreover, we sought to expand the use of the tool from a retrospective perspective (evaluation of existing literature) to a prospective one (planning future studies). This way, researchers will have the opportunity to record all decisions regarding data collection and the consideration of potential confounders in a more systematic way, facilitating the establishment of reporting standards. We decided to focus on acute rather than chronic cortisol level research, which is why we limited the CoAL to cortisol assessment in saliva, blood and urine, and did not include chronic cortisol level assessment methods like hair or fingernail analysis. The end product of this update and extension is the Cortisol Assessment List (CoAL), presented in the current paper.

Our goals for the CoAL can be summarized as follows: First, the CoAL should be both specific and general enough to allow for adaptation to the needs of the users. Second, the CoAL should allow for a dynamic ranking of the items, according to their relative importance in cortisol data collection. Third, the CoAL should meet psychometric standards. In

[1] Reliability is defined as "the trustworthiness or consistency of a measure, that is, the degree to which a test or other measurement instrument is free of random error, yielding the same results across multiple applications to the same sample" [54]. In the context of cortisol assessment, reliability refers to the consistency of cortisol results over time or across participants and the minimization of differences in cortisol concentrations attributable to measurement error. This entails reducing both random and systematic error, for instance by considering potentially confounding influences.

[2] Replication is defined as "the repetition of an original experiment or research study to verify or bolster confidence in its results" [55]. Regarding cortisol assessment, replicability refers to the ability to repeat original studies using the same data collection methodology but different samples with the aim of checking the consistency of the reported results.

practice, these goals can be translated to the following steps:

1) Adapt and extend the items of the Quality Estimation Checklist to create the CoAL
2) Develop a flexible ranking system with regard to the relative importance of the items included in the CoAL
3) Establish the IRR of the CoAL

Steps one and two were realized by conducting a review of the existing literature and conducting an online survey among experts in the field of cortisol assessment. Step three was realized by comparing inter-rater agreement for a set of randomly selected studies assessing cortisol using the CoAL.

2. Methods

In the absence of a broad scientific consensus that addresses the issues raised above, we think that the establishment of a definite checklist with a set ranking system might be premature. We therefore decided to take a first step in creating a comprehensive list of items one could consider for quality ratings or study planning. With such a dynamic list, users still must decide a priori which items are particularly relevant for their specific research endeavors.

2.1. Prerequisites

The CoAL, would have to meet certain general criteria, which we tried to incorporate in the updated design. First, the updated list of items regarding cortisol assessment should be extensive enough to include all the factors that have been shown to influence cortisol levels. However, setting the consideration of all of these factors as the standard for reliable data seems both unrealistic and unnecessary. High quality cortisol data can be obtained by following the majority of recommendations but the decision as to which of them are especially relevant to an individual study may vary according to the respective research question, study design or population of interest.

A second issue related to the selection of items for the list pertains to the lack of a broad scientific consensus regarding which factors are most important to ensure the assessment of reliable cortisol data and which ones could be neglected in favor of a balance between internal and external validity. A list of items that does not differentiate between these "essentials" as opposed to the "nice to haves" may not only be uninformative, it may produce biased estimates of data quality. For example, a study could only consider elements that are easy to implement but are regarded as less essential to maintain high data quality. Vice versa, a study may consider a relatively small number of factors but still obtain reliable data, because they are regarded as essential. A list that statically gives sum scores of the number of items considered in order to determine study quality would not be able to capture this important difference. We tried to tackle these issues by creating a comprehensive and flexibly adaptable list of items for the CoAL.

2.2. Item selection

In line with our goal to reevaluate and extend the Quality Estimation Checklist [28], the existing empirical evidence was reviewed in order to include as many potentially confounding influences as possible. The Quality Estimation Checklist was developed by searching the literature for recommendations on the reporting of sampling procedures and the consideration of relevant confounding influences. It enabled users to evaluate studies measuring cortisol in urine, blood and saliva. Furthermore, it accounted for two different types of study designs: Cortisol assessment could be evaluated in either basal cortisol secretion designs, including CAR and diurnal profile studies, or reactive designs, including psychological, pharmacological or physiological challenge studies. The composition of items in the checklist varied according to these two parameters (specimen and study design). The structure of the original checklist is illustrated in Fig. 1 A. Quality estimates were obtained on four subscales evaluating the *reported sampling design decisions*, the *reported strategies enhancing the accuracy of sampling*, the consideration of variables on the particular day of sampling (*state covariates* [5]; as cited in Ref. [28]) and sociodemographic information (*trait covariates;* [5] as cited in Ref. [28]). In a first step, the individual items of the original checklist were reevaluated. To this end, a second search of the literature was conducted.

We searched the literature for meta research, summarizing the methodological issues that should be considered in order to increase cortisol data quality. Moreover, we rescreened the literature used to create the Quality Estimation Checklist. During this preliminary phase, our intention for the item list was to be as comprehensive as possible. Hence, factors were included as items of the CoAL if the literature suggested an influence in either basal, reactive or CAR designs for urine, blood or saliva sampling. Following extraction, the updated item list was compared to the Quality Estimation Checklist and duplicates were removed before restructuring.

Fig. 1B illustrates the structure of the CoAL. Comparing Fig. 1A and B, one can examine the changes made to the structure of the list. For the CoAL, the general structure of the Quality Estimation Checklist was maintained regarding the specimen of interest, the type of study design and the allocation of the items into various subscales. However, the study design section was further divided to include CAR studies as a separate category, next to basal and reactive cortisol secretion studies. This was done because there was a considerable number of items within the basal cortisol secretion category that pertained to either only studies assessing diurnal cortisol profiles or the CAR. This decrease in overlap prompted us to create separate lists of items in order to maintain a satisfying user experience and avoid confusion.

Another structural change of the CoAL as compared to the Quality Estimation Checklist was the combination of the items in the first two subscales (*reported sampling designs, reported strategies enhancing the accuracy of sampling*) into one subscale called *reporting of sampling procedures.* Lastly, a subscale pertaining to exclusion criteria was added. This resulted in a total of seven possible CoAL compositions, three for cortisol assessment in either blood or saliva (basal cortisol secretion, CAR, reactive cortisol secretion) and one list for cortisol assessment in urine (basal cortisol secretion). The complete lists can be found under https://osf.io/kx3tq/files/.

2.3. Online survey among experts

In line with our goal to provide the users with a preliminary ranking system regarding the relative importance of a certain factor in order to obtain reliable cortisol data, we decided to base this system on the opinions of experts in the field as a first step until a comprehensive expert consensus is established. To this end, we conducted a survey among researchers in the field of psychoneuroendocrinology, asking them to evaluate whether a certain item on the list was in their opinion necessary, desirable or not necessary to ensure high cortisol data quality. Moreover, we asked the respondents to comment on the completeness of the list and make suggestions for items they thought should be included as well.

The survey was advertised during the 2020 virtual annual conference of the International Society for Psychoneuroendocrinology (ISPNE). The rationale behind using a survey approach was to increase retention and get a more systematic overview on the experts' opinions. The ISPNE is the oldest society on psychoneuroendocrinology (PNE) research [29] and its annual conference visitors are mostly experts in the field of psychoneuroendocrinology with years of experience in this line of research. All visitors were invited to take part in the survey. Following the introduction at the conference, additional e-mails with an invitation to participate were sent out to a total of 24 of well-known research groups in the field. Members of these groups were also invited to

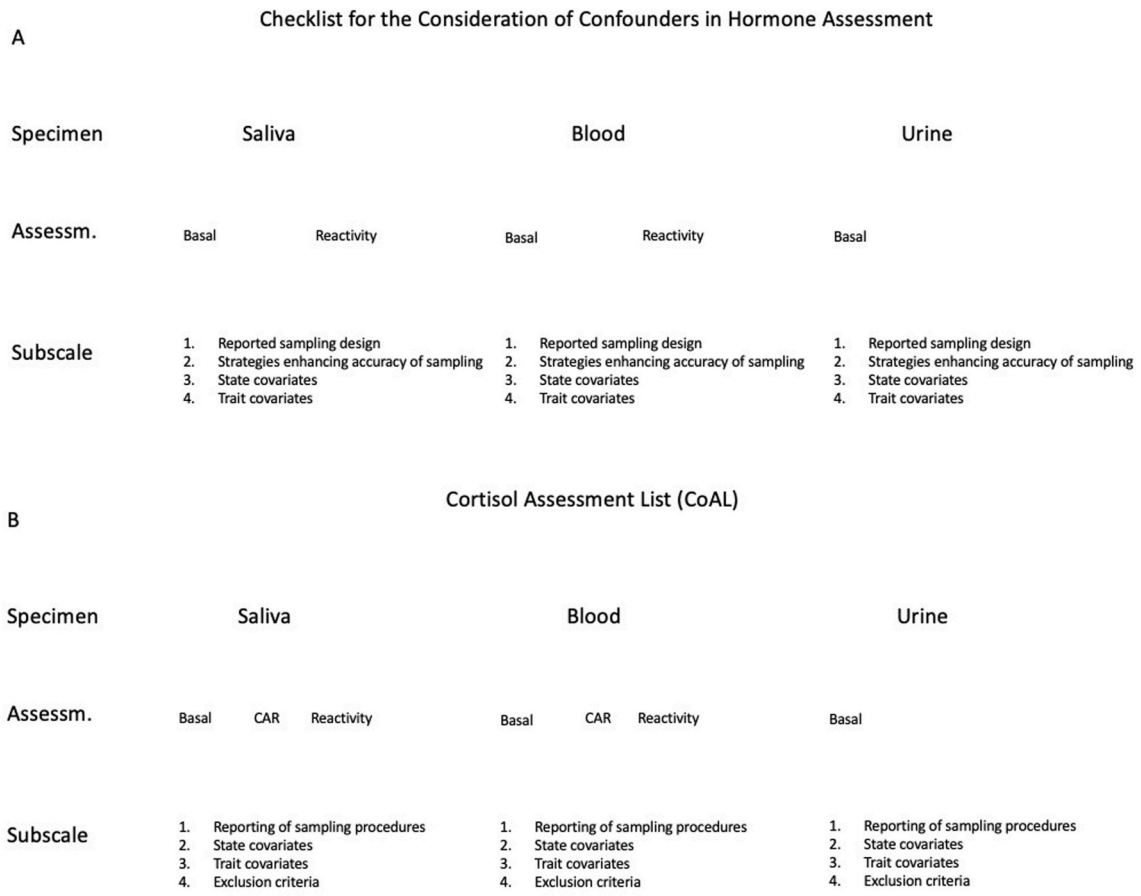

Fig. 1. Changes between the Quality Estimation Checklist and the Cortisol Assessment List (CoAL). A: Structure of the Quality Estimation Checklist. B: Structure of CoAL.

participate in the survey. Data collection took place between September and November 2020.

The online survey was set up using the online survey tool *Unipark* (Questback GmbH, published 2017. EFS Survey, Version Summer 2017. Köln: Questback GmbH). After giving informed consent regarding the goal of data collection, data handling and privacy policy, participants were asked to indicate in which type of study design they had experience (CAR, basal and reactive cortisol secretion designs). Multiple answers were possible.

The survey included four parts with each part representing the items of the four subscales of the CoAL (s.Fig. 1 B) for either CAR, basal or reactive cortisol secretion studies. The questions in each part represented the items to be potentially included in the respective list. Participants were asked to indicate whether they considered an item as *necessary*, *desirable*, or *not necessary* in order to obtain reliable cortisol data. Moreover, participants were able to enter free text under each question. At the end of each part, participants were given the opportunity to comment on items they missed in the list or make other suggestions. At the end of the survey, participants were asked to provide some basic information regarding their academic position, years in the field of psychoneuroendocrinology and the number of published papers.

2.3.1. Survey analysis

Summary statistics were created as the total number of respondents for each of the three answer categories (*necessary*, *desirable*, *not necessary*). Moreover, the free text sections of each item and the comment sections at the end of each page were screened for suggestions that may be of value to the further development of the list.

In order to investigate whether answers differed by level of expertise, we divided the participants according to the number of papers they had published in the field of psychoneuroendocrinology. Participants with zero to ten publications were defined as *junior researchers*, participants with more than 10 publications were defined as *senior researchers*.

Regarding the preliminary classification of items as either *necessary, desirable* or *not necessary* and the suggested addition of new items, a set of rules was defined: Generally, classification decisions were based on the simple majority of votes by senior researchers. For items with an unclear voting result, like two answers receiving the same or close to the same (one vote difference) number of senior votes, the decision was based on the overall majority vote. If this did not resolve the issue and the vote was still inconclusive (i.e., the overall majority vote was contrary to the senior majority vote), items were classified as *desirable*. This was also done for cases in which the majority of the senior vote did not overlap with the overall majority vote. We decided to refrain from putting such items into either the *necessary* or *not necessary* categories, because the intention of the survey was to find the lowest common denominator within the research community, regarding factors that are considered minimal requirements for the reliable assessment of cortisol. Classifying an item with an unclear vote as *necessary* would, in our opinion, not reflect a clear decision towards this goal. Conversely, the potential exclusion of an item from the list without the support of a clear majority of experts seems premature, in absence of a broad scientific consensus.

Regarding suggestions by respondents to add items to the list, the empirical evidence for each suggestion was evaluated for confirmation. In case of inclusion, these items were also put into the *desirable* category as no survey data was available. All commentary including suggestions are available in appendix B.

Following item extraction and integration of the survey data into the CoAL, a user manual, a rating file and an analysis script were created,

which will be described shortly in the following.

2.4. The user manual

The two versions of the user manual provide instructions on how to use the CoAL to either document cortisol assessment in original studies or rate study quality in meta research. The respective user manual contains some general information on the CoAL's intended use and rating decisions which have to be made by the users, followed by a section giving definitions for CAR, active diurnal, and reactive cortisol secretion studies. In the third section of the user manuals, the usage of the CoAL (documentation and quality rating purposes), the rating file and the analysis script (quality rating purposes) are explained in detail. Lastly, the manuals contain a table with explanations on item-level. The user manuals can be found under https://osf.io/kx3tq/files/.

2.5. The rating file

The rating process for the CoAL is identical to the Quality Estimate Checklist. To be more specific, the consideration of a particular potential confounder is indicated by rating the respective item of the CoAL as either *considered* (rate 1), *not considered* (rate 0), or *not applicable* (rate NA). Regarding the definition of the ratings, the users are free to choose, whether they want to pursue a strict approach, rating an item only as *considered*, if it was included in the analyses or controlled for in the study design (only recruiting non-smokers), or a more liberal approach, rating items as *considered* as long as they are assessed during data collection.

The rating file consists of a Microsoft Excel (2021) sheet intended to assist researchers who use the CoAL for quality rating purposes in meta research. It contains one worksheet for each subscale. The items of the CoAL are represented in the columns of the file and studies are rated in the rows. Users can document their rating decisions by entering either the number *1* (considered), *0* (not considered) or the string *NA* (not applicable) in the respective cells. Before starting the rating process, users can determine their individual preference regarding which items to include into the list by entering the word *necessary*, *desirable* or *exclude* in the row of the file called *category*. The default setting of this row reflects the majority vote of the online-survey conducted among experts in the field. The rating file can be downloaded under https://osf.io/kx3tq/files/.

2.6. The analysis script

The analysis script is an R-file (RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/) intended to provide users with the possibility to compute and visualize the percentage of the considered items in the rating file. It takes the rating file as input and calculates percentages for considered items (i.e., those rated with "1") per subscale. Furthermore, an overall score is produced representing the percentage of considered items over all subscales. Results can be visualized as a color-coded tile plot. The analysis script can be downloaded under https://osf.io/kx3tq/files/.

2.7. Psychometric quality determination

In line with our goal to determine the psychometric quality of the CoAL, the inter-rater reliability (IRR) of the CoAL was calculated. To this end, a two-step approach was chosen: In a first step, 75 studies were randomly selected and rated by three pairs of independent raters (25 studies for each pair). The rating team consisted of three students (on Master level) with little experience in cortisol assessment and two of the authors, a doctoral student (SeLa) and a postdoc researcher (SE), who are experienced in conducting studies investigating cortisol. The three rating pairs in this step included one of the three inexperienced student raters and one of the two experienced authors.

Before rating of the 75 selected studies started, the inexperienced raters received general information on cortisol assessment procedures, the different types of study designs and they read the user manual. Subsequently, the rating strategy was discussed, and all raters were instructed to rate one example study. Finally, the ratings of the example studies were discussed between the raters, and last questions resolved.

Of the 75 studies rated in the first step, 25 studies investigated the CAR, basal and reactive cortisol secretion, respectively. Following this first rating process, the IRR for the three rating pairs was calculated and minor adjustments were made to the item selection. These adjustments were based on feedback by the raters and could either entail changes in the composition of items, or their description in the user manual.

In the second step, the adjusted and final list of items was evaluated by rating 15 randomly selected studies with five studies representing each study design (i.e., CAR, basal and reactive cortisol secretion, respectively). The rating pair in this step consisted of two of the authors (SE and SeLa). Ratings were again followed by calculation of the IRR. The study selection process, rating strategy and calculation of the IRR were identical in both rating steps and are outlined in the following.

2.7.1. Study selection

In order to receive the database for the random selection of studies to be rated, a literature search was conducted (performed by SeLa). The database *PubMed* was searched for studies assessing the CAR, active diurnal and reactive cortisol secretion in urine, blood or saliva. Search terms for all three types of study design were entered separately (Appendix A provides a detailed description of the search terms). The literature search was conducted on December 30th' 2020. Filters for the type of publication were: Case reports, classical articles, clinical studies, clinical trials, comparative studies, controlled clinical trials, evaluation studies, journal articles, multicenter studies, pragmatic clinical trials, randomized controlled trials. Publication language was English and publication dates between 1995 and 2020. Publications were then sorted by best match and publication lists were extracted to csv-files.

Using a random number generator (www.random.org), three sets of 25 numbers between one and the respective number of search hits were created. Studies were selected according to their number in the extracted publication list. Inclusion criteria for all studies were: 1) original study, 2) human participants, 3) cortisol assessment matches the search term (CAR, basal cortisol secretion, reactive cortisol secretion, see Appendix A for details). Following the preliminary selection of 75 studies, the abstracts and methods sections of all papers were screened according to the inclusion criteria. In case one of the selected studies did not match the inclusion criteria, the publication listed above in the publication list was selected. Following the sorting, this publication was the next best match according to the search criteria entered in the online search. In case of another exclusion, this process was repeated until a study matching the inclusion criteria was found. If the selected study referred to another paper for the detailed description of sample collection, this study was selected as a replacement. A complete list of the studies selected for rating can be found in appendix A.

2.7.2. Rating strategy

Regarding the rating strategy, we chose a rather liberal approach. An item was rated as considered, if it was either statistically controlled for or if the publication reported that it was measured. This approach was chosen, because the focus at this point was the determination of agreement between raters, rather than a stricter estimation of cortisol assessment quality or a quantification of considered factors in cortisol assessment studies.

2.7.3. Determination of inter-rater reliability

The IRR for the rater pairs across the respective studies were calculated using R software (RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/). To this end, the ratings of each pair were

compared for each study. Mismatches were investigated and disagreement between raters was resolved through discussion [30]. In case of a clear rating error by one of the raters (i.e. the information could (not) be found back in the paper), rating decisions were corrected. This was also done, if there was a systematic misunderstanding regarding rating decisions. Following this, the IRR was calculated as Cohen's Kappa [31].

3. Results

3.1. Online survey

3.1.1. Participants

In total, 27 participants started the survey of whom two did not complete, leaving a total of 25 completers. Of the 25 respondents who completed the online survey, 12 were categorized as *seniors* (>10 publications), the remaining 13 were categorized as *juniors*. According to their specific areas of expertise, 13 respondents (8 *seniors*, 5 *juniors*) completed the CAR section, 12 (8 *seniors*, 4 *juniors*) the active (diurnal) cortisol secretion section and 21 (11 *seniors*, 10 *juniors*) the reactive cortisol secretion section. The academic positions held by the respondents were master's student (n = 1), PhD candidate (n = 5), postdoc (n = 10) and professor (n = 8). One respondent stated to be a private lecturer. The average amount of years of experience in cortisol assessment was 12.48 (9.13) with *seniors* having significantly more experience than *juniors* (*seniors*: M = 20.25, SD = 6.70; *juniors*: M = 5.31, SD = 3.07; $t = 7.265$, $p < 0.001$).

3.1.2. Survey voting results

The survey results are summarized in Table 1 (a more detailed version on item-level is provided in appendix B). The ratio of items that were considered *necessary* and *desirable* was quite consistent across all three study design sections, with slightly more items being considered *necessary* by the majority of experts. The only clear deviation from this was found for state covariates in reactive cortisol secretion designs (6 items *necessary*, 10 items *desirable*). Only very few items were deemed *not necessary* (k = 2 for CAR; k = 1 for active diurnal cortisol secretion; k = 1 for reactive cortisol secretion). These items included the coupling of intravenous sampling with polysomnography for CAR studies and the consideration of birth weight as a trait covariate for all three study design sections. We decided to remove the item related to the coupling of intravenous sampling with polysomnography but leave the consideration of birth weight in. This approach was chosen as the literature suggests that birth weight may have an impact in both active diurnal and reactive cortisol secretion [32–34]. Moreover, birth weight is a variable widely considered in cortisol research involving infants or smaller children. This decision is in line with the overall goal to pursue maximal universality for the CoAL, irrespective of study population or study design. Nevertheless, we labeled birth weight as *not necessary* in the default setting of the respective rating files.

Conversely, the literature regarding coupling of intravenous cortisol sampling to polysomnography is sparse and there is no clear evidence suggesting a superior benefit of coupling blood collection to polysomnographic activity. The study mentioned in the CAR guidelines by Stalder and colleagues (2016), timed forced awakening to automatic sample collection ([35] as cited in Ref. [5]). However, we could not find any studies on a direct comparison between this methodology and other collection methods to capture the CAR. Hence, the item was removed from the list.

There were a few items for which the vote was indecisive or there was a discrepancy between the majority vote of seniors compared to the overall majority vote. As mentioned before, these items were classified as *desirable*. Table 2 gives an overview of the items in question and the respective distribution of votes.

3.1.3. Free text entries

Lastly, the free text fields were screened regarding suggestions for items to be added to the lists. This led to the addition of three items, which were classified as *desirable* due to the absence of a vote. The first item suggested to be added was *acute alcohol consumption on the day of sampling* (a state covariate, as opposed to habitual alcohol consumption, which we classified as a trait covariate). Subsequently, the evidence and suggestions of earlier publications were screened again. Strahler et al. [6] suggest controlling for both acute and habitual alcohol consumption, even though it is stated that the empirical evidence on the influence of acute alcohol consumption on cortisol is somewhat inconclusive. However, it was decided to follow this suggestion, which is in line with our main goal of creating a comprehensive list.

The second item added upon suggestion *was recreational drug use*. This item was added to the subscale for exclusion criteria. The effects of recreational drug use on the HPA axis are well documented for a variety of drugs such as marihuana [36–39], opioids [40,41], or cocaine [42, 43]. The empirical evidence strongly suggests considering this during the sample selection process.

The third item added upon analysis of the survey data was *mental disorders*. In light of the evidence regarding disorders like depression, anxiety disorders or PTSD [8,44–46], this item was also added to the exclusion criteria list.

A last item suggested for addition to the list was *athlete status* of the sample. It was decided to not add this item to the list at this point. There is one study finding an effect of athlete status on hair cortisol concentrations [47], but the evidence regarding cortisol levels in blood, saliva or urine rather points towards a negligible influence [6,48].

3.2. Study ratings

Further adjustments to the CoAL were implemented after the first round of study ratings. The items *accounting for time of day of sampling* and *relating sampling to clock/wake up time* were merged into the item *fixed sampling time points either related to clock times or wake up time*. Moreover, the items *accounting for the menstrual cycle phase* and *accounting for menopause* were merged into *accounting for menstrual cycle phase or menopause*. The item *use of a diary log reporting exact sampling and wake up time* was split up into *use of a diary log reporting exact sampling times* and *use of a diary log reporting exact wake up time.* The items *reminding participants the night before sampling, sampling time points reported in paper,* and *experiences on the day prior to sampling* were deleted. Lastly, the wording of four items was changed. A detailed overview of all these changes can be found in appendix C.

These changes resulted in the final sets of items, comprising the CoAL. There are seven combinations of items according to the type of

Table 1
Number of items classified as *necessary*, *desirable* or *not necessary*, according to the majority vote.

CoAL module	Number of items			
CAR	Necessary	Desirable	Not necessary	Total
Reporting of design procedures	6	3	1	10
State covariates	9	7	0	16
Trait covariates	6	5	1	12
Total	21	15	2	40
Basal cortisol secretion				
Reporting of design procedures	6	5	0	11
State covariates	7	8	0	15
Trait covariates	6	5	1	12
Total	20	17	1	38
Reactive cortisol secretion				
Reporting of design procedures	5	3	0	8
State covariates	6	10	0	16
Trait covariates	6	5	1	12
Total	17	18	1	36
Exclusion criteria	7	0	0	7

Note: The columns on the right show the number of items for each of the experts in the field of psychoneuroendocrinology.

Table 2
Survey items with an indecisive voting result.

Section	Text	Vote	
Subcategory		Necessary	Desirable
CAR	Objectively monitoring sampling	Total: 6	Total: 7
Reported design procedures	time (time-stamped containers etc.)	Sr.: 5	Sr.: 3
		Jr.: 1	Jr.: 4
CAR	Menopause	Total: 5	Total: 7
Trait covariates		Sr.: 4	Sr.: 3
		Jr.: 1	Jr.: 4
CAR	Ethnicity	Total: 4	Total: 6
Trait covariates		Sr.: 4	Sr.: 2
		Jr.: 0	Jr.: 4
CAR	Perceived chronic stress	Total: 4	Total: 8
Trait covariates		Sr.: 4	Sr.: 3
		Jr.: 0	Jr.: 5
Basal cortisol secretion	Brushing teeth	Total: 5	Total: 6
State covariates		Sr.: 4	Sr.: 3
		Jr.: 1	Jr.: 3
Basal cortisol secretion	Menopause	Total: 5	Total: 6
Trait covariates		Sr.: 4	Sr.: 3
		Jr.: 1	Jr.: 3
Reactive cortisol secretion	Precise description of potential habituation effects	Total: 8	Total: 11
Reported design procedures		Sr.: 6	Sr.: 4
		Jr.: 2	Jr.: 7
Reactive cortisol secretion	Time of awakening	Total: 10	Total: 8
State covariates		Sr.: 2	Sr.: 6
		Jr.: 8	Jr.: 2
Reactive cortisol secretion	Day of sampling (weekday vs. weekend)	Total: 9	Total: 8
State covariates		Sr.: 3	Sr.: 6
		Jr.: 6	Jr.: 2
Reactive cortisol secretion	Habitual alcohol intake (drinks per week)	Total: 11	Total: 9
Trait covariates		Sr.: 4	Sr.: 6
		Jr.: 7	Jr.: 3
Reactive cortisol secretion	Season of the year	Total: 9	Total: 5
State covariates		Sr.: 3	Sr.: 4
		Jr.: 6	Jr.: 1
Reactive cortisol secretion	Anticipation of the day load	Total: 12	Total: 8
State covariates		Sr.: 4	Sr.: 7
		Jr.: 8	Jr.: 1
Reactive cortisol secretion	Childhood adversity	Total: 12	Total: 7
Trait covariates		Sr.: 3	Sr.: 6
		Jr.: 9	Jr.: 1

Note: **Sr.** = Senior researchers with more than ten publications in the field of psychoneuroendocrinology. **Jr.** = Junior researchers with less than ten publications in the field of psychoneuroendocrinology.

cortisol assessment (CAR, active diurnal or reactive cortisol assessment) and the specimen collected (blood, saliva, or urine). Accordingly, the number of items varies between 29 (Active diurnal cortisol assessment in urine) and 48 (Active diurnal cortisol assessment in saliva). The complete lists, rating files and analysis scripts are available for download under https://osf.io/kx3tq/files/.

3.3. Inter-rater reliability

The IRR for the first round of ratings expressed as Cohen's Kappa was k = .94 for studies assessing the CAR, k = .97 for studies assessing active diurnal cortisol secretion and k = .96 for studies assessing reactive cortisol secretion, indicating very high agreement. For the second round, the IRR was k = .98 (CAR); k = .99 (active diurnal); k = .99 (reactive) respectively, which is in line with the first round.

3.4. Example ratings

Again, it should be noted that the primary goal of the current rating approach was not to determine study quality but rather the agreement between different raters. Hence, the results of the quality ratings are not discussed in detail. However, in order to illustrate how the CoAL ratings can be analyzed and presented (in systematic reviews), we created a tile plot showing the percentages of items rated as *considered* for the five studies rated in the second rating round (Fig. 2A). This tile plot was created using the r-software analysis script which is also available for download (https://osf.io/kx3tq/files/; RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com). The percentages of considered items are displayed per subscale and in an overall score. Note that this tile plot used all the items in the CoAL as input irrespective of the voting results obtained from the online survey among experts. To illustrate the impact of different item weighting approaches, which can be specified individually, according to each researcher's preferences, we created a second tile plot, including only the items that were rated as *necessary* by the majority of experts in the online survey (Fig. 2B). As can be seen, there is a clear change in the percentages for all subscales, indicating that most of the items rated as *considered* for the five studies seem to have been items that the majority of experts regarded as necessary to obtain reliable cortisol data.

4. Discussion

The CoAL is to our knowledge the first comprehensive list of factors that, if considered, increase the reliable assessment of cortisol in saliva, blood or urine. The CoAL is an easily accessible, transparent and standardized tool to document cortisol assessment in original studies. Furthermore, the CoAL was developed as a tool to systematically evaluate cortisol assessment in meta research. Lastly, the CoAL meets psychometric standards and strikes a balance between theory and practice regarding the importance of the respective items in order to obtain reliable cortisol data which is supported by the majority of the scientific community. In accordance with the Open Science Principles, all the materials (i.e. the CoAL, the rating file and the analysis script) are freely available online, helping to improve transparency in cortisol assessment.

The results of the survey among experts in the field helped significantly in approaching this balance. The classification of items on the CoAL as *necessary, desirable,* or *not necessary* was quite consistent across all 25 respondents. Our results are a valid representation of the current opinion in the field, regarding what constitutes the lowest common denominator in enabling reliable cortisol assessment. This claim is supported by the high degree of expertise in our sample with 19 of 25 respondents holding postdoc or professorial positions, and an average work experience of 12 years in the field of psychoneuroendocrinology. Moreover, it should be noted that all of the items added to the CoAL are based on empirical evidence showing an influence on cortisol levels. Therefore, it might be assumed that the majority of the respondents has had to decide whether or not to consider these factors at some point during data collection for their empirical research. One could thus infer that most of the respondents were aware of the empirical evidence for or against the *necessity* of a certain item in order to obtain reliable cortisol data. This can also be observed by the relatively few *not sure* responses and the extensive written comments indicating either support or opposition to the respective items. The few items regarded as *not necessary* by the majority of our sample also indicate high awareness of the potential influences of all factors included in the CoAL.

Regarding the determination of psychometrics for the CoAL, the high interrater reliability (IRR) shown for a broad variety of randomly selected studies suggests that it fulfills the psychometric requirements possible for a tool of this kind. Furthermore, inter-rater agreement was high for both, raters with high and low levels of experience in cortisol assessment, showing that the CoAL is well suited to be used by

Fig. 2. Example of rating results for the five randomly selected studies assessing the CAR. Numbers in the tile-plot represent the percentage of items rated as *considered*. **A**: Even weighting over all items included in the CoAL. **B**: Weighting according to the majority vote of the online survey, including only items that were considered *necessary* by the majority of respondents.

researchers with varying degrees of expertise.

In summary, both the online survey and the study rating results show that the CoAL is a versatile, comprehensive and well-balanced tool to evaluate and document a wide variety of cortisol assessment designs in saliva, blood or urine.

4.1. Practical implications for the CoAL

The CoAL enables researchers to make a conscious decision for or against the consideration of a potential factor of influence during the planning phase, and before data collection has started. This reduces the probability of unintentional flaws that are only discovered after the fact.

The accessible documentation of all cortisol assessment decisions in a standardized list, also increases replicability. In practice, this could entail including the CoAL in the appendices of original studies, which provides the opportunity to replicate cortisol collection procedures a lot more precisely, as some details may not be stated explicitly in publications due to word count constraints. In our opinion, inter-study comparability is a key factor in gaining significant insights in any field of research. Additionally, inconclusive results might be traceable a lot better as a result of this increase in transparency.

Next to documentation, the CoAL can also be used to evaluate cortisol assessment quality according to a set of pre-defined standards, enabling meta research endeavors to systematically document the consideration of assessment procedures known to influence the reliability of cortisol levels. A practical implication is the transparent documentation of what constitutes high data quality in advance of meta-analytical data collection and the weighting of study results. This way, a compromise between strict inclusion criteria at the cost of study quantity and loose inclusion criteria at the cost of data quality can be achieved, while maintaining sufficient flexibility for an individual definition of high-quality cortisol assessment. To be more specific, researchers can adapt the list to their specific needs if their rationale for choosing a particular set of items is reported. This again increases transparency and may lead to the disclosure of research gaps for future endeavors.

Lastly, the CoAL may assist in peer review processes, especially for reviewers who are not familiar with cortisol assessment conventions and may wish to have a handy tool to get a rough estimate regarding the quality of the cortisol data presented in a paper.

However, the practical implications of the CoAL also warrant a clear illustration regarding issues that are currently unresolved. The absence of comprehensive cortisol assessment guidelines points out the first major issue regarding the use of the CoAL: Generalizability across meta research implementing the CoAL.

While the high adaptability regarding the weight of items used to determine data quality expands applicability, it certainly limits between-study comparability. In absence of an expert consensus, we do not want to preempt the discussion and set fixed criteria for the CoAL. On the one hand, this enables users to apply potential future cortisol assessment guidelines to the CoAL, once they are established, and still use the CoAL in the meanwhile. On the other hand, the CoAL does not provide a standardized and comparable score at the moment. More specifically, quality estimates depend on the items selected for the CoAL. Hence, statements regarding study quality are only applicable to the item composition used for the respective meta-analysis. This of course limits the use of the CoAL as an "objective" measure of cortisol assessment quality. Even more so, the configuration of items may even be misused in order to create weightings that favor a certain meta-analytical outcome. To prevent misuse, we strongly recommend to pre-register any specifications made with regard to the CoAL (i.e., item composition and weightings) and report item configuration transparently. The default configuration of the CoAL, as indicated by the online survey results, offers a good starting point and deviations from these settings (i.e., leaving items off the CoAL that are considered *necessary*) should be justified. If users adhere to these recommendations, the CoAL is in our opinion a very useful tool for meta research.

A second unresolved issue of the CoAL in its current form is that it cannot give an estimate of what constitutes a *good* study. The number of factors that are documented during the planning phase (prospective use of the CoAL) or the scores obtained through the rating process (retrospective use of the CoAL) reflect the percentage of items rated as *considered*. However, without a fixed set of items, it is in our view premature to determine which percentage of considered items constitutes a study for which sufficient cortisol data quality can be assumed. A precondition for such a determination is the aforementioned agreement among the scientific community regarding the minimum requirements to ensure high quality data. Nevertheless, we think the mere percentage of considered items can provide valuable information and as long as the criteria are reported transparently, we do not see a problem in letting researchers decide for themselves, where they want to set cut-offs.

The third issue regarding the practical implementation of the CoAL is that it does at this point not include cortisol assessment in hair or fingernails. We specifically did not include the assessment of cortisol in hair or fingernails since cortisol concentrations in these specimens reflect HPA axis activity over a longer time frame (up to several weeks) as compared to the other three assessment methods, reflecting short-term HPA axis activity (usually over several days at most). Hence, the potentially confounding influences and respective recommendations in hair cortisol tend to differ significantly from the other assessment methods, especially regarding the specifics of sample collection (for an overview see Refs. [49–52]). Research on confounders in fingernail

cortisol assessment is mainly still being conducted at the moment, and has to our knowledge been reviewed [53], but not been summarized on a meta-analytical level, yet. We encourage all interested parties to extend the CoAL to the documentation and evaluation of hair and fingernail cortisol.

Lastly, an inherent practical challenge of the CoAL must be mentioned. Since there is no comparable tool assessing cortisol data quality or an objective measure against which the CoAL could be compared, we cannot make any claims regarding the validity of the CoAL. Hence, it cannot be said with certainty that the CoAL truly serves as an estimate of cortisol data quality. Furthermore, the CoAL has not been used for documentation purposes in original studies yet. For this reason, we are currently only able to make claims regarding its reliability in meta research. However, all the items included in the CoAL are based on empirical evidence and our survey findings further support the notion that the CoAL includes all factors that are considered to be the most important in order to obtain reliable cortisol data. Future research may of course identify new ways to increase reliability and should certainly evaluate the CoAL's construct validity in documenting high quality cortisol assessment once more data is available. We consider the CoAL to be a work in progress that can be expanded according to current state of knowledge. This requires regular updates, which are planned in the future.

To conclude, the practical implementation in original studies and meta research is pivotal to future adaptations and updates to the CoAL and the current version must be understood as a first step on the road to comprehensive guidelines for cortisol assessment.

4.2. Limitations

In addition to the currently unresolved issues regarding the practical application of the CoAL, some limitations regarding the study should be listed. Firstly, the expert sample recruited may only represent a specific part of the scientific community in PNE research. We advertised the survey at the annual conference of the ISPNE, which is the oldest and largest society for PNE research. Nevertheless, the survey may not have reached researchers with expertise in reliable cortisol assessment if they either did not attend this conference or were not made aware of the survey while attending.

We tried to counteract this bias in sample selection by contacting 24 well known laboratories with an invitation to participate in the survey and encouraged these laboratories to spread the information within their list of contacts. Our intention was to spread awareness of the survey and not limit participants to attendees of the conference. Still, we recognize, that our final sample may be composed of a specific field of researchers at one of many conferences on cortisol research. This again highlights the urgent need to establish a body of researchers who represent all varieties of cortisol research in order to establish comprehensive assessment guidelines that extend the already published guidelines introduced for the CAR [5].

A second limitation related to the sample selection is our method of creating a proxy estimate of an expert consensus, based on the results of an online survey. We are aware, that an online survey cannot replace the process of a true broad scientific consensus, which often involves several opportunities for discussion and an in-depth study of the empirical evidence by all parties involved. This was obviously not the case here. Hence, it is important to note that the survey should only be viewed as a first impression of the opinions of experts in the field. This is especially important for all items with unclear voting results, which we classified as *desirable* as to not jump ahead in any potential discussions to be held in the future. The adaptable item weighting system of the CoAL enables the users to take this issue into account, allowing for adaptations where need be. Nevertheless, this flexibility of the weighting system means that the current item weights in the CoAL should be revisited and, if necessary, adapted with regard to the users' specific research questions and the populations of interest.

A third limitation of this study is that we did not define 'reliability' in the introductory text of the survey, so there may be some variability to what researchers considered this term to entail. We assumed all respondents to have a comprehensive understanding of what reliability means in cortisol research (i.e. the minimization of systematic and random measurement error in order to enable replicable data) and that these definitions would greatly overlap. Still, we cannot rule out that some participants did not have such a comprehensive understanding of the term, which may have led to some variance.

5. Conclusion and future directions

As mentioned before, the CoAL can serve well as a starting point for any discussion around cortisol assessment guidelines based on a broad scientific consensus among cortisol researchers. Furthermore, it could serve as a 'mirror' reflecting current scientific practices, once original studies include it in their appendices. More specifically, future research may summarize the consideration frequencies of the CoAL items in original studies. This could inform consensus talks regarding the question what is ideal and what is practical.

In case of adoption as a documentation tool of cortisol assessment in blood, saliva and urine, it is conceivable to extend the CoAL to hair and fingernail cortisol assessment or even other hormones. The ultimate goal should be the establishment of assessment and reporting standards in the whole field of psychoneuroendocrinology.

In conclusion, the CoAL reflects empirical evidence and expert knowledge regarding cortisol assessment practices. It meets psychometric standards and can be used as a flexible tool to report empirical studies or evaluate cortisol data quality in meta research. In line with the Open Science initiative, it is freely available, and we aim to provide a valuable tool to the scientific community.

Author contribution form (CRediT)

Sebastian Laufer: Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, Writing – original draft. Sinha Engel: Formal analysis, investigation, methodology, writing – review & editing. Sonia Lupien: Supervision, writing – review & editing. Christine Knaevelsrud: Project administration, supervision, writing – review & editing. Sarah Schumacher: Conceptualization, methodology, project administration, supervision, writing – review & editing.

Declaration of competing interest and funding

This research did not receive any specific grant from funding sources in the public, commercial, or not-for-profit sectors. Declarations of interests: none.

Acknowledgements

The authors would like to thank Mrs. Annika Montag and Mrs. Jana Funke for their contribution in rating the studies and summarizing the survey results. A special thank you to Dr. Lars Schulze for his valuable contributions in the creation of the analysis script and helpful comments regarding data visualization. Lastly, we would like to thank all of the survey participants who provided us with their expertise in the field of cortisol assessment.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cpnec.2021.100108.

References

[1] E.K. Adam, M. Kumari, Assessing salivary cortisol in large-scale, epidemiological research, Psychoneuroendocrinology 34 (10) (2009) 1423–1436, https://doi.org/10.1016/j.psyneuen.2009.06.011.

[2] K. Hanrahan, A.M. McCarthy, C. Kleiber, S. Lutgendorf, E. Tsalikian, Strategies for salivary cortisol collection and analysis in research with children, Appl. Nurs. Res. 19 (2) (2006) 95–101, https://doi.org/10.1016/j.apnr.2006.02.001.

[3] B.M. Kudielka, D.H. Hellhammer, S. Wüst, Why do we respond so differently? Reviewing determinants of human salivary cortisol responses to challenge, Psychoneuroendocrinology 34 (1) (2009) 2–18, https://doi.org/10.1016/j.psyneuen.2008.10.004.

[4] R. Ryan, S. Booth, A. Spathis, S. Mollart, A. Clow, Use of salivary diurnal cortisol as an outcome measure in randomised controlled trials: a systematic review, Ann. Behav. Med.: Publ. Soc. Behav. Med. 50 (2) (2016) 210–236, https://doi.org/10.1007/s12160-015-9753-9.

[5] T. Stalder, C. Kirschbaum, B.M. Kudielka, E.K. Adam, J.C. Pruessner, S. Wüst, S. Dockray, N. Smyth, P. Evans, D.H. Hellhammer, R. Miller, M.A. Wetherell, S. J. Lupien, A. Clow, Assessment of the cortisol awakening response: expert consensus guidelines, Psychoneuroendocrinology 63 (2016) 414–432, https://doi.org/10.1016/j.psyneuen.2015.10.010.

[6] J. Strahler, N. Skoluda, M.B. Kappert, U.M. Nater, Simultaneous measurement of salivary cortisol and alpha-amylase: application and recommendations, Neurosci. Biobehav. Rev. 83 (2017) 657–677, https://doi.org/10.1016/j.neubiorev.2017.08.015.

[7] S. Laufer, S. Engel, C. Knaevelsrud, S. Schumacher, Cortisol and alpha-amylase assessment in psychotherapeutic intervention studies: a systematic review, Neurosci. Biobehav. Rev. 95 (2018) 235–262, https://doi.org/10.1016/j.neubiorev.2018.09.023.

[8] S. Schumacher, H. Niemeyer, S. Engel, J.C. Cwik, S. Laufer, H. Klusmann, C. Knaevelsrud, HPA axis regulation in posttraumatic stress disorder: a meta-analysis focusing on potential moderators, Neurosci. Biobehav. Rev. 100 (2019) 35–57, https://doi.org/10.1016/j.neubiorev.2019.02.005.

[9] N.F. Narvaez Linares, V. Charron, A.J. Ouimet, P.R. Labelle, H. Plamondon, A systematic review of the Trier Social Stress Test methodology: issues in promoting study comparison and replicable research, Neurobiol. Stress 13 (2020) 100235, https://doi.org/10.1016/j.ynstr.2020.100235.

[10] S. Fischer, J.M. Doerr, J. Strahler, R. Mewes, K. Thieme, U.M. Nater, Stress exacerbates pain in the everyday lives of women with fibromyalgia syndrome—the role of cortisol and alpha-amylase, Psychoneuroendocrinology 63 (2016) 68–77, https://doi.org/10.1016/j.psyneuen.2015.09.018.

[11] M. Meier, L. Wirz, P. Dickinson, J.C. Pruessner, Laughter yoga reduces the cortisol response to acute stress in healthy individuals, Stress 24 (1) (2021) 44–52, https://doi.org/10.1080/10253890.2020.1766018.

[12] J. Hellhammer, E. Fries, O.W. Schweisthal, W. Schlotz, A.A. Stone, D. Hagemann, Several daily measurements are necessary to reliably assess the cortisol rise after awakening: state- and trait components, Psychoneuroendocrinology 32 (1) (2007) 80–86, https://doi.org/10.1016/j.psyneuen.2006.10.005.

[13] B.M. Kudielka, L.C. Hawkley, E.K. Adam, J.T. Cacioppo, Compliance with ambulatory saliva sampling in the chicago health, aging, and social relations study and associations with social support, Ann. Behav. Med.: Publ. Soc. Behav. Med. 34 (2) (2007) 209–216, https://doi.org/10.1007/BF02872675.

[14] J.E. Broderick, D. Arnold, B.M. Kudielka, C. Kirschbaum, Salivary cortisol sampling compliance: comparison of patients and healthy volunteers, Psychoneuroendocrinology 29 (5) (2004) 636–650, https://doi.org/10.1016/S0306-4530(03)00093-3.

[15] V.C. Smith, L.R. Dougherty, Noisy spit: parental noncompliance with child salivary cortisol sampling, Dev. Psychobiol. 56 (4) (2014) 647–656, https://doi.org/10.1002/dev.21133.

[16] N. Jacobs, N.A. Nicolson, C. Derom, P. Delespaul, J. van Os, I. Myin-Germeys, Electronic monitoring of salivary cortisol sampling compliance in daily life, Life Sci. 76 (21) (2005) 2431–2443, https://doi.org/10.1016/j.lfs.2004.10.045.

[17] M.A. Babyak, Understanding confounding and mediation, Evid. Base Ment. Health 12 (3) (2009) 68–71, https://doi.org/10.1136/ebmh.12.3.68.

[18] A. DeMaris, Combating unmeasured confounding in cross-sectional studies: evaluating instrumental-variable and Heckman selection models, Psychol. Methods 19 (3) (2014) 380–397, https://doi.org/10.1037/a0037416.

[19] K.A. Frank, Impact of a confounding variable on a regression coefficient, Socio. Methods Res. 29 (2) (2000) 147–194, https://doi.org/10.1177/0049124100029002001.

[20] L.T. Hoyt, K.B. Ehrlich, H. Cham, E.K. Adam, Balancing scientific accuracy and participant burden: testing the impact of sampling intensity on diurnal cortisol indices, Stress 19 (5) (2016) 476–485, https://doi.org/10.1080/10253890.2016.1206884.

[21] S.H. Golden, B.N. Sánchez, M. Wu, S. Champaneri, A.V. Diez Roux, T. Seeman, G. S. Wand, Relationship between the cortisol awakening response and other features of the diurnal cortisol rhythm: the Multi-Ethnic Study of Atherosclerosis, Psychoneuroendocrinology 38 (11) (2013) 2720–2728, https://doi.org/10.1016/j.psyneuen.2013.06.032.

[22] D.L. Hall, D. Blyler, D. Allen, M.H. Mishel, J. Crandell, B.B. Germino, L.S. Porter, Predictors and patterns of participant adherence to a cortisol collection protocol, Psychoneuroendocrinology 36 (4) (2011) 540–546, https://doi.org/10.1016/j.psyneuen.2010.08.008.

[23] K. Valentino, A. De Alba, L.C. Hibel, K. Fondren, C.G. McDonnell, Adherence to diurnal cortisol sampling among mother–child dyads from maltreating and nonmaltreating families, Child. Maltreat. 22 (4) (2017) 286–294, https://doi.org/10.1177/1077559517725208.

[24] M. Stoffel, A.B. Neubauer, B. Ditzen, How to assess and interpret everyday life salivary cortisol measures: a tutorial on practical and statistical considerations, Psychoneuroendocrinology 133 (2021) 105391, https://doi.org/10.1016/j.psyneuen.2021.105391.

[25] J.L. Calvi, F.R. Chen, V.B. Benson, E. Brindle, M. Bristow, A. De, S. Entringer, H. Findlay, C. Heim, E.A. Hodges, H. Klawitter, S. Lupien, H.M. Rus, J. Tiemensma, S. Verlezza, C.-D. Walker, D.A. Granger, Measurement of cortisol in saliva: a comparison of measurement error within and between international academic-research laboratories, BMC Res. Notes 10 (1) (2017) 479, https://doi.org/10.1186/s13104-017-2805-4.

[26] W.K. Goodman, J. Janson, J.M. Wolf, Meta-analytical assessment of the effects of protocol variations on cortisol responses to the Trier Social Stress Test, Psychoneuroendocrinology 80 (2017) 26–35, https://doi.org/10.1016/j.psyneuen.2017.02.030.

[27] J.M. Hulett, J.M. Armer, E. Leary, B.R. Stewart, R. McDaniel, K. Smith, R. Millspaugh, J. Millspaugh, Religiousness, spirituality, and salivary cortisol in breast cancer survivorship: a pilot study, Cancer Nurs. 41 (2) (2018) 166–175, https://doi.org/10.1097/NCC.0000000000000471.

[28] S. Schumacher, H. Niemeyer, S. Engel, J.C. Cwik, C. Knaevelsrud, Psychotherapeutic treatment and HPA axis regulation in posttraumatic stress disorder: a systematic review and meta-analysis, Psychoneuroendocrinology 98 (2018) 186–201, https://doi.org/10.1016/j.psyneuen.2018.08.006.

[29] F. Brambilla, 1969–1989: twenty years of life of the international society of psychoneuroendocrinology, Psychoneuroendocrinology 14 (4) (1989) 247–249, https://doi.org/10.1016/0306-4530(89)90028-0.

[30] M. Borenstein (Ed.), Introduction to Meta-Analysis, John Wiley & Sons, 2009.

[31] R.G. Orwin, Evaluating coding decisions, in: The Handbook of Research Synthesis, Russell Sage Foundation, 1994, pp. 139–162.

[32] D.I.W. Phillips, B.R. Walker, R.M. Reynolds, D.E.H. Flanagan, P.J. Wood, C. Osmond, D.J.P. Barker, C.B. Whorwood, Low birth weight predicts elevated plasma cortisol concentrations in adults from 3 populations, Hypertension 35 (6) (2000) 1301–1306, https://doi.org/10.1161/01.HYP.35.6.1301.

[33] R.M. Reynolds, B.R. Walker, H.E. Syddall, R. Andrew, P.J. Wood, C.B. Whorwood, D.I.W. Phillips, Altered control of cortisol secretion in adult men with low birth weight and cardiovascular risk factors 86 (1) (2001) 6.

[34] S. Wüst, S. Entringer, I.S. Federenko, W. Schlotz, D.H. Hellhammer, Birth weight is associated with salivary cortisol responses to psychosocial stress in adult life, Psychoneuroendocrinology 30 (6) (2005) 591–598, https://doi.org/10.1016/j.psyneuen.2005.01.008.

[35] I. Wilhelm, J. Born, B.M. Kudielka, W. Schlotz, S. Wüst, Is the cortisol awakening rise a response to awakening? Psychoneuroendocrinology 32 (4) (2007) 358–366, https://doi.org/10.1016/j.psyneuen.2007.01.008.

[36] T.T. Brown, A.S. Dobs, Endocrine effects of marijuana, J. Clin. Pharmacol. 42 (S1) (2002) 90S–96S, https://doi.org/10.1002/j.1552-4604.2002.tb06008.x.

[37] E.J. Cone, R.E. Johnson, J.D. Moore, J.D. Roache, Acute effects of smoking marijuana on hormones, subjective effects and performance in male human subjects, Pharmacol. Biochem. Behav. 24 (6) (1986) 1749–1754, https://doi.org/10.1016/0091-3057(86)90515-0.

[38] A. Cservenka, S. Lahanas, J. Dotson-Bossert, Marijuana use and hypothalamic-pituitary-adrenal Axis functioning in humans, Front. Psychiatr. 9 (2018) 472, https://doi.org/10.3389/fpsyt.2018.00472.

[39] M. Ranganathan, G. Braley, B. Pittman, T. Cooper, E. Perry, J. Krystal, D. C. D'Souza, The effects of cannabinoids on serum cortisol and prolactin in humans, Psychopharmacology 203 (4) (2009) 737–744, https://doi.org/10.1007/s00213-008-1422-2.

[40] A. Fountas, S.T. Chai, C. Kourkouti, N. Karavitaki, Mechanisms OF endocrinology: endocrinology of opioids, Eur. J. Endocrinol. 179 (4) (2018) R183–R196, https://doi.org/10.1530/EJE-18-0270.

[41] M. Walter, G.A. Wiesbeck, B. Degen, J. Albrich, M. Oppel, A. Schulz, H. Schächinger, K.M. Dürsteler-MacFarland, Heroin reduces startle and cortisol response in opioid-maintained heroin-dependent patients: heroin reduces startle and cortisol response, Addiction Biol. 16 (1) (2011) 145–151, https://doi.org/10.1111/j.1369-1600.2010.00205.x.

[42] H.C. Fox, E.D. Jackson, R. Sinha, Elevated cortisol and learning and memory deficits in cocaine dependent individuals: relationship to relapse outcomes, Psychoneuroendocrinology 34 (8) (2009) 1198–1207, https://doi.org/10.1016/j.psyneuen.2009.03.007.

[43] C.M. Heesch, B.H. Negus, R.W. Snyder, E.J. Eichhorn, J.H. Keffer, R.C. Risser, Effects of cocaine on cortisol secretion in humans, Am. J. Med. Sci. 310 (2) (1995) 61–64, https://doi.org/10.1097/00000441-199508000-00004.

[44] G.P. Chrousos, Stress and disorders of the stress system, Nat. Rev. Endocrinol. 5 (7) (2009) 374–381, https://doi.org/10.1038/nrendo.2009.106.

[45] C.M. Pariante, S.L. Lightman, The HPA axis in major depression: classical theories and new developments, Trends Neurosci. 31 (9) (2008) 464–468, https://doi.org/10.1016/j.tins.2008.06.006.

[46] G.E. Tafet, C.B. Nemeroff, The links between stress and depression: psychoneuroendocrinological, genetic, and environmental interactions, J. Neuropsychiatry Clin. Neurosci. 28 (2) (2016) 77–88, https://doi.org/10.1176/appi.neuropsych.15030053.

[47] N. Skoluda, L. Dettenborn, T. Stalder, C. Kirschbaum, Elevated hair cortisol concentrations in endurance athletes, Psychoneuroendocrinology 37 (5) (2012) 611–617, https://doi.org/10.1016/j.psyneuen.2011.09.001.

[48] T. Cevada, P. Vasques, H. Moraes, A. Deslandes, Salivary cortisol levels in athletes and nonathletes: a systematic review, Horm. Metab. Res. 46 (13) (2014) 905–910, https://doi.org/10.1055/s-0034-1387797.
[49] E. Russell, G. Koren, M. Rieder, S. Van Uum, Hair cortisol as a biological marker of chronic stress: current status, future directions and unanswered questions, Psychoneuroendocrinology 37 (5) (2012) 589–601, https://doi.org/10.1016/j.psyneuen.2011.09.009.
[50] T. Stalder, S. Steudte-Schmiedgen, N. Alexander, T. Klucken, A. Vater, S. Wichmann, C. Kirschbaum, R. Miller, Stress-related and basic determinants of hair cortisol in humans: a meta-analysis, Psychoneuroendocrinology 77 (2017) 261–274, https://doi.org/10.1016/j.psyneuen.2016.12.017.
[51] T. Stalder, C. Kirschbaum, Analysis of cortisol in hair – state of the art and future directions, Brain Behav. Immun. 26 (7) (2012) 1019–1029, https://doi.org/10.1016/j.bbi.2012.02.002.
[52] S.M. Staufenbiel, B.W.J.H. Penninx, Y.B. de Rijke, E.L.T. van den Akker, E.F.C. van Rossum, Determinants of hair cortisol and hair cortisone concentrations in adults, Psychoneuroendocrinology 60 (2015) 182–194, https://doi.org/10.1016/j.psyneuen.2015.06.011.
[53] S. Fischer, S. Schumacher, N. Skoluda, J. Strahler, Fingernail cortisol – state of research and future directions, Front. Neuroendocrinol. 58 (2020) 100855, https://doi.org/10.1016/j.yfrne.2020.100855.
[54] American Psychological Association, APA Dictionary of Psychology (n.d.), https://dictionary.apa.org/reliability.
[55] American Psychological Association, APA Dictionary of Psychology (n.d.), https://dictionary.apa.org/replication.