Simple, Transparent, and Flexible Automated Quality Assessment Procedures for Ambulatory Electrodermal Activity Data

Ian R. Kleckner*, Rebecca M. Jones, Oliver Wilder-Smith, Jolie B. Wormwood, Murat Akcakaya, Karen S. Quigley, Catherine Lord, and Matthew S. Goodwin

Abstract*— *Objective: **Electrodermal activity (EDA) is a non-invasive measure of sympathetic activation often used to study emotions, decision-making, and health. The use of "ambulatory" EDA in everyday life presents novel challenges—frequent artifacts and long recordings—with inconsistent methods available for efficiently and accurately assessing data quality. We developed and validated a simple, transparent, flexible, and automated quality assessment procedure for ambulatory EDA data.** ***Methods:*** **Twenty individuals with autism (5 females, 5-13 years) provided a combined 181 hours of EDA data in their home using the Affectiva Q Sensor across 8 weeks. Our procedure identified invalid data using four rules: (1) EDA out of range; (2) EDA changes too quickly; (3) temperature suggests the sensor is not being worn; and (4) transitional data surrounding segments identified as invalid via the preceding rules. We identified invalid portions of a pseudo-random subset of our data (32.8 hours, 18%) using our automated procedure and independent visual inspection by five EDA experts.** ***Results:*** **Our automated procedure identified 420 minutes (21%) of invalid data. The five experts agreed strongly with each other (agreement: 98%, Cohen's *κ:* 0.87) and thus were averaged into a "consensus" rating. Our procedure exhibited excellent agreement with the consensus rating (sensitivity: 91%, specificity: 99%, accuracy: 92%, *κ:* 0.739 [95% CI=0.738, 0.740]).** ***Conclusion:*** **We developed a simple, transparent, flexible, and automated quality assessment procedure for ambulatory EDA data.** ***Significance:*** **Our procedure can be used beyond this study to enhance efficiency, transparency, and reproducibility of EDA analyses, with free software available at http://www.cbslab.org/EDAQA.**

***Index Terms*—Electrodermal activity, data quality assessment, quality control, wearables**

I. Introduction

Electrodermal activity (EDA) is a non-invasive peripheral measure of sympathetic nervous system activation commonly used to assess physiological arousal [2]. EDA is typically measured using a recording device containing two small sensors placed on the skin of the fingers, palm, feet, or other parts of the body [3, 4]. The sensors complete a circuit passing through the skin (a resistor) and the EDA recording device measures the fluctuations in conductance of the skin due to changes in the amount of sweat in eccrine gland ducts, which are controlled by the sympathetic nervous system. EDA is conventionally decomposed into background or tonic skin conductance level (SCL), which encompasses relatively slow and continuous changes in EDA over tens of seconds, and skin conductance responses (SCRs), which are relatively fast and discrete events superimposed on the SCL (by convention, an SCR is typically considered an increase in EDA of at least 0.05 µS over 1-3 sec followed by a slower decrease in EDA toward its pre-SCR level over 3-15 sec) [4-6]. However, EDA is not a perfect measure of sympathetic nervous system activation because eccrine sweat gland density differs across sites on the body [3], EDA does not necessarily reflect sympathetic activation to other organs in the periphery [7], and EDA is affected by non-sympathetic factors such as environmental temperature and humidity [2].

Despite these limitations, EDA can be recorded easily and non-invasively and thus has been used extensively to study physiological arousal in emotion [8], attention [9], decision-making [10], pain [11], stress [12], autism [13], phobias [14], panic disorder [15], attention deficit disorders [16], side-effects

This research was supported in part by the Simons Foundation (336363 to M.S.G., I.R.K., C.L. and R.M.J), the National Institute of Nursing Research (NR013500 to M.S.G.), the National Institute on Deafness and Other Communication Disorders (P50 DC013027 to M.S.G.), the National Cancer Institute (R25 CA102618 and UG1 CA189961 to support I.R.K.), and a grant from the U.S. Army Research Institute for the Behavioral and Social Sciences (W911NF-16-1-0191 to K.S.Q. and J.B.W). The views, opinions, and/or findings contained in this paper are those of the authors and shall not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

I.R.K. is with the University of Rochester Medical Center, Rochester, NY. USA.

R.M.J and C.L are with Weill Cornell Medicine, The Center for Autism and the Developing Brain, White Plains, NY, USA.

O.W.S., J.B.W., and M.S.G. are with Northeastern University, Boston, MA, USA.

K.S.Q. is with the Edith Nourse Rogers Memorial (VA) Medical Center, Bedford, MA and Northeastern University, Boston, MA

M.A. is with the University of Pittsburgh, Pittsburgh, PA, USA.

* Corresponding author: Ian_Kleckner@URMC.Rochester.edu

from cancer treatments [17], and other psychological phenomena (for a review, see [2]). Traditionally, studies using EDA have been performed in carefully controlled laboratory settings with brief recording durations (often a few minutes per participant) with procedures to minimize participant movement and variations in temperature and humidity, and utilizing pre-selected, time-locked stimuli or tasks. Recent advances in mobile sensing technology [18] have enabled broader use of "ambulatory" EDA in everyday life [19]. Ambulatory EDA studies provide enormous amounts of data—e.g., tens of hours per participant—and capture open-ended naturalistic settings such as life at home, school, work, or during stressful events. The advances afforded by ambulatory EDA have enhanced ecological validity by revealing whether laboratory-based phenomena also occur in naturalistic settings, and provided a window into phenomena that do not occur in controlled laboratory environments [18], such as individuals with flight phobia flying in an airplane [14]. More generally, recording EDA in both laboratory and naturalistic settings enables one to explain environment-specific variance.

Analysis of ambulatory EDA data requires tremendous care given the unique challenges it presents—such as frequent artifacts and varied data quality throughout long recordings—which occur due to greater movement in daily life. Recording artifacts can arise from a variety of sources, such as: (1) movement artifacts from pressure or movement of the electrodes relative to the skin [20]; (2) participants intentionally or unintentionally touching the sensors (especially for individuals with sensory sensitivity conditions such as autism); or (3) contextual factors (e.g., air humidity, temperature) that cause excessive sweating that increases the EDA signal beyond a device's capabilities (i.e., "bridging," "ceiling effect," "saturation") [1]. These issues are further compounded by the fact that ambulatory EDA recordings are typically extremely long, sometimes with tens of hours per participant. Whereas traditional quality assessment of laboratory-based EDA data can be performed by rigorous and methodical visual inspection and human coding, this time-consuming process does not scale well to ambulatory EDA datasets. Although the technology to acquire ambulatory EDA data has increased rapidly, the development of software to perform automated quality assessment on ambulatory EDA data has not kept pace.

There is a critical need for simple, transparent, and flexible automated quality assessment procedures for ambulatory EDA data. There are several extant software programs to help detect and analyze individual SCRs (e.g., MindWare EDA Program, Biopac AcqKnowledge, [21-23]), and emerging software programs to help distinguish SCRs from artifacts (e.g., [24, 25]). However, these analytic tools were developed for laboratory-acquired EDA data, typically recorded under conditions of little to no participant movement. Fewer tools are available for automated quality assessment of ambulatory EDA data beyond identification and analysis of SCRs, and existing tools have only recently begun to appear in the literature (e.g., [26-28]). These tools, although automated and somewhat flexible, are still being optimized and are neither simple nor entirely transparent, making independent interpretation and replication by others difficult. Simplicity is desirable (although not essential) to help researchers and readers gain a mutual understanding of the methods used so they can be evaluated, built upon, and easily adapted to match recording parameters of different devices. If a technique is not transparent, it can hinder reproducibility and lead to undetected artifacts that can influence results and subsequent interpretations. For example, in recent work by Taylor et al. [28], automated EDA quality assessments were made using a machine learning algorithm to distinguish "valid" vs. "invalid" 5-sec segments of EDA data according to ratings by two human EDA experts. The machine learning algorithm had access to 14 different EDA features (mean EDA, maximum slope, etc.) and performed well (96% accuracy) in replicating decisions of human EDA experts when the two experts agreed with each other (although performance suffered when EDA experts disagreed on EDA data validity). Although sophisticated, the authors did not indicate what rules the machine learning model used to identify any portion of EDA data as "valid" or "invalid" based on a complex 14-dimensional EDA feature space. This problem is significant for researchers trying to understand *why* certain portions of their EDA data are invalid (e.g., to provide better instructions to researchers and participants to improve data quality, for building a sensor that is more robust to specific types of artifacts). The lack of transparency is also problematic for replicating findings. In the current world of "big data"—including longitudinal ambulatory EDA studies—the need for transparent automated methods has never been greater. Indeed, the National Institutes of Health (NIH) recently released an initiative designed to support greater rigor and reproducibility in biomedical research [29].

The goal of the current study was to develop and evaluate a simple, transparent, and flexible automated quality assessment procedure for ambulatory EDA data. The contributions of this work include: (1) development of an automated EDA quality assessment procedure using four simple and adjustable rules; (2) demonstration that results from our automated procedure agree with results from the human EDA raters; and (3) demonstration of better agreement with human raters than an extant alternative approach by Taylor et al. [28]. We tested our automated procedure on a large ambulatory EDA dataset acquired in the home from 20 children and adolescents with autism. We used home-based data because it provided a wide range of EDA features without knowledge of context—where EDA quality decisions had to be made based on EDA data alone—which was precisely the condition for which we designed our procedure. Next, to obtain quality assessment of the same EDA data, five independent human EDA raters (all with substantial prior expertise evaluating EDA data quality) identified invalid portions of data by visual inspection of raw EDA signals. We evaluated our automated procedure by comparing its performance to that of five human raters and to that of Taylor et al. [28]. Our free, open-source software is available at http://www.cbslab.org/EDAQA.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2758643, IEEE Transactions on Biomedical Engineering

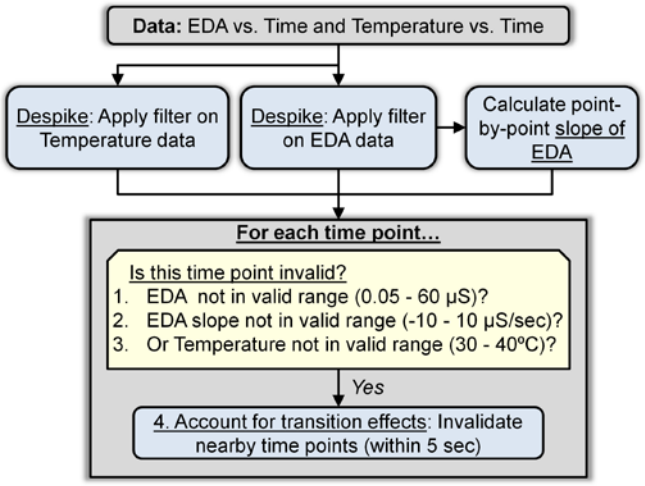

Fig. 1. Our automated quality assessment procedure uses a despiking filter and four simple rules to determine if each data point is invalid or valid. The values in parentheses were used for this study but can be modified based on differences in EDA hardware, participant sample, experimental context, and the relative importance of keeping vs. removing both valid and invalid data.

II. Materials and Methods

A. Participants.

Twenty children and adolescents with a confirmed diagnosis of autism spectrum disorder from a licensed clinician at the Center for Autism and the Developing Brain (CADB) and their families were recruited to participate (5 females, age range 5-13 years, mean age = 8 years). All caregivers provided informed consent and children and adolescents seven years of age and older provided assent in accordance with the Institutional Review Board at Weill Cornell Medicine.

B. Procedures

The participants and their caregivers participated for a total of eight weeks. There was a total of three clinic visits to CADB during weeks 1, 4, and 8. During the first clinic visit, caregivers were trained how to operate the Q Sensor (Affectiva, Inc.; Waltham, MA) and were provided with detailed instructions for turning on/off the device, synchronizing the sensor clock to Universal Time, placing electrodes on their child's wrist to optimize signal quality, and charging the device after each use. The Q Sensor acquired EDA, temperature, and 3-axis accelerometer data at a sampling frequency of either 16 Hz or 32 Hz. To enhance sensitivity, solid conductive adhesive hydrogel Ag/AgCl electrodes were used (22 mm square; model A10040-5 from Vermed; Buffalo, NY). The Q Sensor was positioned on the ventral surface (underside) of the non-dominant wrist. Caregivers were provided with an athletic sweatband to place over the Q Sensor to prevent the child from touching it, mistakenly turning the device off, or moving the electrodes. At home, caregivers were instructed to put the Q Sensor on their child for three separate days during the week following their clinic visit for approximately 1.5 hours per day. Caregivers were told the child should go about their daily activities as usual, but to remove the device before bathing or showering. Caregivers were also given written instructions on how to operate the device, and a research assistant periodically checked in with the family to address any issues.

In addition to the Q Sensor, caregivers were provided with a mobile application on their smartphone to provide daily feedback about their child's behavior (mood, irritability, and disruptive behaviors) and the children wore a Language Environment Analysis (LENA) device to record their spoken language. Data from these measures are beyond the scope of this study and will be presented elsewhere.

TABLE 1
OUR FOUR SIMPLE RULES FOR QUALITY ASSESSMENT OF EDA DATA

Rule	Rationale
1. EDA is out of range (not within 0.05-60 µS)	To prevent "floor" artifacts (e.g., electrode loses contact with skin) and "ceiling" artifacts (circuit is overloaded). We chose 0.05 µS because it is at an accepted minimum for SCR amplitude [1]. We chose 60 µS because it is near the maximum for the Q Sensor.
2. EDA changes too quickly (faster than ±10 µS/sec)	To prevent high frequency or "jump" artifacts [2] (e.g., [4, 6]).
3. Temperature is out of range (not within 30-40ºC)	To account for times when the EDA sensor is not being worn or has not been worn long enough. Our data reached plateaus of approximately 32-36ºC across individuals, and in this temperature window, we also achieved our most stable electrode-skin interface as evidenced by stable EDA measures.
4. EDA data are surrounding (within 5 sec of) invalid portions according to rules 1-3	To account for transition effects close in time to artifacts.

C. Developing the automated quality assessment procedure

The in-home data consisted of 181 hours of EDA data across 195 recordings. Fig. 1 shows a flow chart of our automated quality assessment procedure. First, we used a low-pass "despiking" filter to remove noise, as recommended by [2]. Specifically, we used a low-pass FIR filter of length 2,057 (symmetric around zero) with cutoff frequencies at 0 and 0.35 Hz and designed for a sampling frequency of 32 Hz. Because the filter is FIR and symmetric, the phase response is linear and this corresponds to a constant group delay of 1,028 samples. We selected this filter to capture the (low frequency) changes in SCL and potential SCRs, which vary from approximately 3-15 sec. After filtering, we tested whether each data point was valid or invalid based on the four rules shown in Table 1.

Data marked invalid by these four rules are not likely to be associated with any physiological abnormalities or pathology. Specifically, whereas certain conditions such as clinical anxiety have been associated with modest increases in EDA level or EDA slope at rest and in response to laboratory stimuli [2], these changes are small compared to the physiologically unrealistic values indicated by the rules in Table 1.

D. Quality assessment by human EDA experts

To test the performance of our automated procedure against assessments of EDA experts, we pseudo-randomly selected 100 approximately 20-min-long segments of data. The procedure selected a 20-min segment starting at least 10 min into a recording (to ensure that stable contact between the Q Sensor

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2758643, IEEE Transactions on Biomedical Engineering

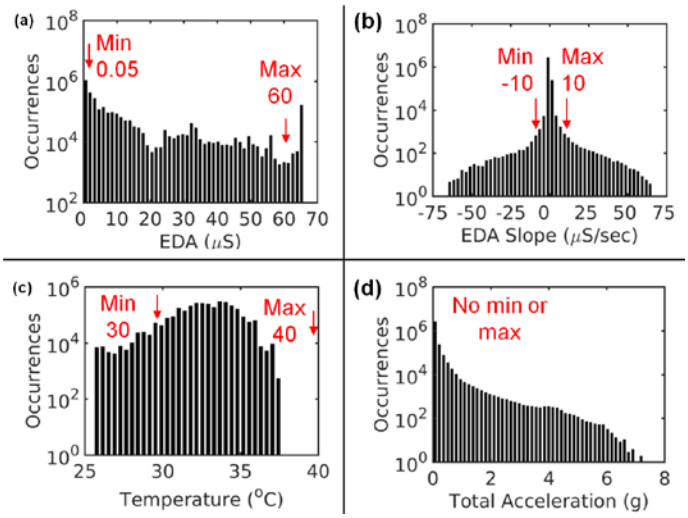

Fig. 2. Distributions of data features across all data points in all 100 files with minimum and maximum allowable values used in our quality assessment procedure marked in red. (a) EDA level (b) EDA slope, (c) temperature, and (d) total acceleration, the latter of which was not used in our procedure.

and the skin was established) from a random subset of files, where each file was a single recording session from a specific participant and recording day. We started with files at least 30 min long, and obtained all 94, 20-min segments; we then chose 6 more files and extracted segments ranging from 11.7 to 19.2 min. The total duration of data for automated and human quality assessment was 1,972.3 min (average file duration of 19.7 min).

Five raters (I.R.K., O.W.S., J.B.W., K.S.Q., M.S.G.), all with extensive expertise conducting EDA analyses, each provided independent manual ratings of all 1,972.3 min (100 files) of EDA data. The raters were instructed to use only the rules indicated in Table 1 as visual heuristics (not mathematical calculations), with the exception that the human raters did not view the temperature data. Raters used web-based software that we developed to visualize EDA data to select and mark invalid segments [30]. The software loads and displays EDA data, permits zoom in and out functionality, allows the user to highlight and add custom text labels to portions of EDA data, and allows the user can revisit any file at any time. Human raters were instructed to identify invalid data to the nearest second. The human raters also provided confidence ratings (high or low) for their selection of each invalid segment.

Some datasets were removed because they could not be rated by some EDA experts due to an error in the web-based EDA visualization software. This occurred for only 22 dataset-rater combinations (4.4% of total data). Due to the small amount of data affected, we do not expect this to meaningfully bias our results. One data file had no ratings from any human EDA expert and thus was removed. In total, 90% of the data received ratings from all five EDA experts, and there was at least one human EDA expert rating for each data point in 1,952.3 min of data (spanning 99 of the 100 files).

E. Comparing EDA ratings within raters and between raters and our automated procedure

We concatenated the data across all 99 datasets within each EDA rater to create a vector where each element indicates whether the data point was valid or invalid based on its rating (each data point reflects 1/32 sec or 1/64 sec of data, based on the sampling frequency). We compared raters to each other in a pairwise manner (10 pairwise comparisons) using percent agreement and Cohen's κ [31], and then averaged the 10 values of percent agreement and κ. We also created a human consensus EDA expert based on majority vote to mark each data point as either valid or invalid.

We compared results from the human consensus rating to the results of our automated procedure to yield percent agreement, sensitivity (number of time points where both human consensus and the automated procedure indicated "valid" divided by the number of time points that human consensus indicated "valid"), specificity (number of time points where both human consensus and the automated procedure indicated "invalid" divided by the number of time points that human consensus indicated "invalid"), and κ. All calculations were performed using MATLAB (MathWorks; Natick, MA).

F. EDA quality assessment using EDA Explorer by Taylor et al. [28]

We analyzed our 100 datasts with the web-baesd softare from Taylor et al. [28]. We accessed the website, uploaded data, and completed analyses on May 8, 2017 using the version of the software available at the time with its default settings. Results were returned in 5-sec epochs, and thus comparisons between Taylor, our automated procedure, and human experts were all made in 5-sec epochs that each reflect majority vote (valid or invalid) across the time points in the epoch.

III. Results

A. Distributions of data features

Fig. 2 shows different features of our Q Sensor dataset including EDA level, EDA slope, temperature, and acceleration. Across all data points in the 100 data files, there was significant variability in all Q Sensor features. The EDA level was typically in the lower range of the scale (median 2.8 μS, 95% CI = 0, 65 μS), although a significant portion of data was outside the allowable range (0.05-60 μS). For EDA slope, the clear majority of data were within the allowable range of ±10 μS/sec (median 0 μS/sec, 95% CI -0.6, 0.6 μS/sec), but there were some points with very rapid changes in EDA. The temperature data were often 33-35ºC (median 32.9ºC, 95% CI = 28.7, 35.8ºC), reflecting that the Q Sensor was worn for at least 10 min and no more than 30 min. Finally, the acceleration data, like the EDA data, were typically in the lower range of the scale (median 0.06 g, 95% CI = 0.003, 0.52 g) with few occasions of high acceleration. This suggests that participants were more often sedentary or only moderately active during recording periods, although such designations are imperfect using only wrist-worn accelerometers [32].

B. Results from our automated procedure

Fig. 3 shows five examples of invalid portions of data automatically identified by our procedure. The examples include violations of Rule 1 (EDA is out of range; Fig. 3a-b), Rule 2 (EDA changes too fast; Fig. 3c), Rule 3 (temperature is out of range; Fig. 3e), and Rule 4 set to mark as invalid a 5-sec

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2758643, IEEE Transactions on Biomedical Engineering

range of data around invalid points from rules 1-3 (Fig. 3d). These examples show that our procedure successfully removed invalid regions of data, although sometimes at the expense of also removing valid regions of data, such as in Fig. 3e where some rapid EDA changes were removed (red arrow at 13 min). Across all 1,972.3 min of data (100 files), our automated procedure identified 420 min (21.3%) of invalid data.

C. *Results from human EDA raters*

Analogous to our automated procedure, five human EDA raters independently identified invalid regions of EDA data across all 99 approximately 20-min files using our web-based software tool to view, zoom in/out, and highlight invalid data based on violations of our rules concerning EDA level (not within 0.05-60 µS) and EDA slope (changing faster than ±10 µS/sec). Each human rater required approximately 2 hours to evaluate all 99 files. Across all raters, 58% of the invalid segments were identified with high confidence and the remaining 42% were identified with low confidence. Two of the raters always identified invalid segments with low confidence, and the other three raters identified 46%, 50%, or 86% of invalid segments with high confidence. Despite differences in confidence, we found excellent agreement among the five raters, with average inter-rater agreement of 98% and average inter-rater κ of 0.87 (ranging 0.82-0.94 across the 10 pairs of raters, with all ps $< 10^{-90}$). This high level of agreement justified creating a consensus EDA rating by majority vote (valid or invalid) across all raters for each data point.

D. *Comparison of our automated procedure and the consensus EDA rater*

Qualitatively, there was excellent agreement between automated and human quality consensus assessment procedures. However, Fig. 4 shows a few examples of discrepancies. For instance, the red arrow in Fig. 4a shows that the automated procedure marked data as invalid due to low temperature, whereas the human consensus rater marked the data as valid; the blue arrow in Fig. 4d shows that the automated procedure marked the data as valid whereas the human consensus rater marked the data as invalid due to the large, rapid spikes. Quantitatively, our automated procedure exhibited excellent agreement with the consensus rating with sensitivity of 91%, specificity of 99%, accuracy of 92%, and κ of 0.739 (95% CI = 0.738, 0.740, $p < 10^{-90}$) across three million data points (Table 2). The table shows that most errors from our automated procedure were "misses," i.e., the automated procedure marked data as invalid in cases where human EDA consensus marked data as valid. Thus, with these EDA data and rules, our automated procedure is relatively liberal in excluding data; that is, our procedure removes most invalid data but sometimes also removes valid data.

E. *Comparison of our automated procedure to the automated procedure of Taylor et al.*

We also ran our 100 datasets through the web-based algorithm of Taylor, et al. [28]. The two approaches exhibited only modest agreement, with overall agreement of 73% and κ of 0.39 (95% CI = 0.38, 0.41, $p < 10^{-90}$). The Taylor et al. procedure was more likely to mark data as invalid (38% invalid) than our current approach (24% invalid; n.b., the slight difference of 24% invalid here vs. 21% invalid from Section B is because this comparison required decimating our data to Taylor's 5-sec sampling period). Most of the time (55% of data points), both methods indicated the data point was valid, and some of the time (18%) both methods indicated the data point was invalid. In cases where the approaches disagreed, 20% of the time only our procedure indicated the data point was valid; in rarer instances (6%), only the Taylor procedure indicated the

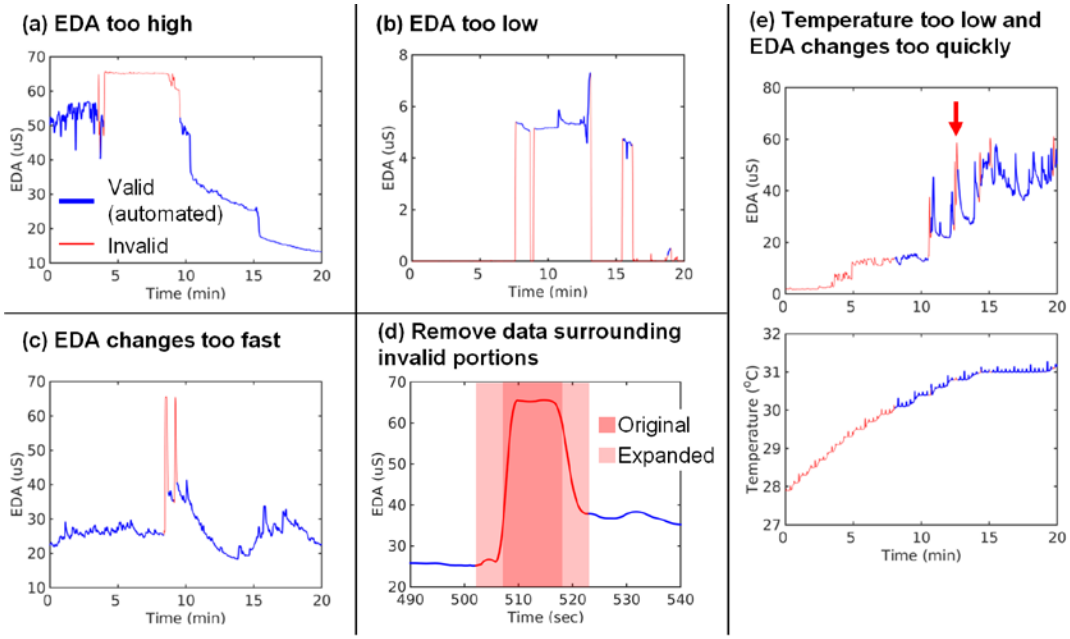

Fig. 3. Examples of invalid data assessed by our automated procedure due to each EDA rule: (1) EDA range set to 0.05-60 µS; (2) EDA maximum slope set to ±10 µS/sec; (3) temperature range set to 30-40°C; and (4) transition range set to any data within 5 sec. Valid data are blue and invalid data are red. In panel (d), the dark red shaded portion shows data originally removed because EDA level and slope were too high. The light red shaded portion shows that region expanded to a 10-sec wide window (Rule 4 in Table 1) to account for transition effects surrounding the artifact. In panel (e), the red arrow indicates a region where the automated procedure removed data that appears to be valid. The data shown were subject to our low-pass filter.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2758643, IEEE Transactions on Biomedical Engineering

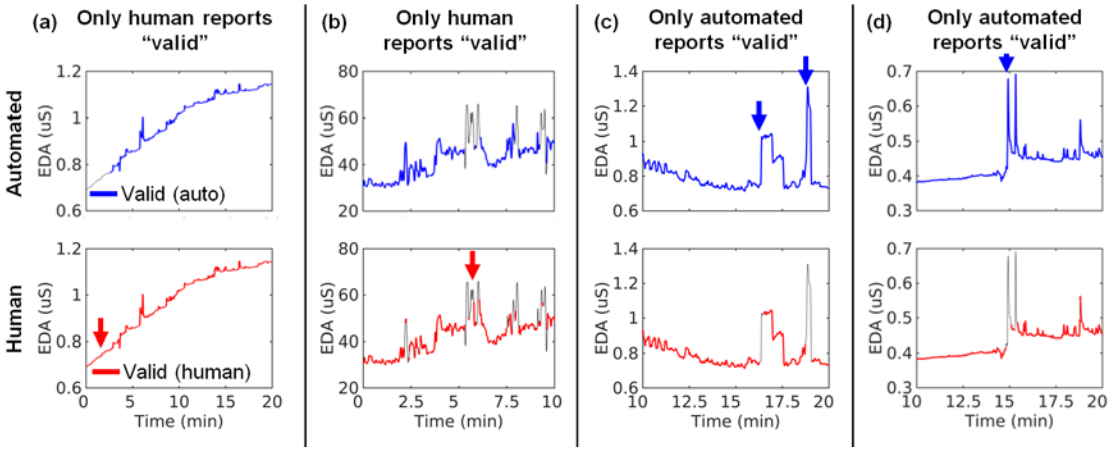

Fig. 4. Example discrepancies between our automated procedure (top row) and human EDA experts (bottom row). The red arrows indicate areas where only a human EDA expert indicated the data were valid. The blue arrows indicate areas where only the automated procedure indicated the data were valid. The data shown were subject to our low-pass filter.

data point was invalid. Finally, the Taylor procedure exhibited poorer agreement with the human consensus rating than did the current procedure, with overall agreement of 77% and κ of 0.46 (95% CI = 0.45, 0.48). By comparison, our procedure exhibited an overall agreement of 94% and κ of 0.83 with the human consensus rating (presented in section D).

IV. Discussion

This study presents a simple, transparent, and flexible automated quality control procedure for ambulatory EDA data that exhibits excellent agreement with human EDA experts. This procedure is particularly relevant given the need for rigorous, transparent, and reproducible science—per the NIH's recent initiative [17]—and the need for automation in analyses of ambulatory EDA data, which are typically too large for complete visual inspection by human raters. Using data from individuals with autism, a particularly challenging population from which to obtain high quality EDA data, we were able to demonstrate the validity of our procedure in recording conditions that were far less optimal than a laboratory setting with compliant, unmoving, typically developing adults.

It is important to note that researchers using our procedure will need to adjust the four rules based on their recording device specifications and study design. Parameter choices emphasize the inherent trade-off between *keeping* vs. *removing* both valid and invalid data. We are not suggesting canonical rules for the field to follow regardless of data properties because rigid rules would not be appropriate across different devices, contexts, and study goals. Instead, we recommend that researchers use an independent set of data to establish criteria for the automated procedure that match manual ratings. To avoid bias, quality assessment criteria should be established before examining results for the primary EDA-based outcome of the study. We also recommend that authors provide their quality assessment specifications to help to establish useful criteria for future studies.

Our work takes a complementary approach to that of Taylor et al. [28], which distinguished between invalid and valid EDA data segments by training a support vector machine classifier on 14 EDA features (e.g., mean, maximum value of first temporal derivative). Whereas a support vector machine classifier that makes use of many different features obscures the rules being implemented in determining valid vs. invalid EDA data, our approach uses four transparent rules, making it easy to understand how it operates. Additionally, the Taylor et al. study exhibited relatively low agreement between the two human EDA experts (81% agreement and Cohen's κ of 0.55); this erodes validity because it is not clear if the algorithm was trained against a reliable standard. By comparison, our work obtained ratings from five EDA experts and we achieved high inter-rater agreement (average 98% inter-rater agreement and average inter-rater κ of 0.90). In one of the analyses of Taylor et al., they also trained their classifier to distinguish three classes of EDA data: clean (both raters indicated data were clean), questionable (only one of the two raters indicated the data were clean), and artifact (both raters indicated the data were artifactual). By comparison, we used only two classes of data, consistent with our observation of very high inter-rater agreement. Finally, Taylor et al.'s datasets contained more invalid data (artifacts) than the current work (39% vs. 21% invalid), which might explain some of these differences in performance. However, when we applied the Taylor algorithm to our dataset, we found that it did not perform as well as our approach, when both were compared to the human raters. Even with further fine-tuning to improve the Taylor approach to our data, it is unlikely to significantly surpass the excellent agreement that our procedure obtained with human EDA experts, which approached the ceiling (sensitivity of 91%, specificity of 99%, accuracy of 92%).

TABLE 2
COMPARISON OF AUTOMATED PROCEDURE AND HUMAN CONSENSUS EDA EXPERT

	Hits	Misses	False Alarms	Correct Rejections	Overall Performance			
					Accuracy	Sensitivity	Specificity	Cohen's κ
Human Consensus EDA Expert	Valid	Valid	Invalid	Invalid	92%	91%	99%	0.739
Automated Procedure	Valid	Invalid	Valid	Invalid				95% CI = 0.738, 0.740
Time Points	2,488,095	236,719	6,663	440,421				

It is also important to consider the strengths and limitations of using a rule-based algorithm (presented here) compared to a machine-learning based algorithm, such as a support vector machine for EDA quality control (e.g., Taylor et al. [28]). The strengths of a rule-based algorithm include transparency and simplicity, making it easy for researchers to see which rules are most useful and which rules have caused undesirable results. By extension, the lack of transparency in machine learning algorithms is its primary weakness. When machine learning models become very complex, it may be prohibitively difficult to determine which features cause the model to make poor decisions. Moreover, machine learning approaches have a greater risk of over-fitting data, causing poor generalizability to situations that are not similar enough to the training data. The weaknesses of a rule-based algorithm are that it might not fully exploit all features of the data, and thus cannot match human performance in many areas unless those features are directly coded as rules. It is also arguable whether researchers can identify all the relevant rules in EDA quality control, thus placing an upper-limit on the performance of rule-based approaches. However, our data suggest excellent performance using a rule-based algorithm in the current dataset. In addition, certain quality control decisions might involve combinations of features, not just simple rules as we propose. Indeed, there may be unknown features associated with EDA data that human raters anchor on in making quality control decisions. By extension, the strengths of a machine learning approach are that it can exploit many features simultaneously to more closely match human performance. Future research should continue to consider a wide range of methods to optimize automated quality assessment of EDA data. In addition, rules to perform quality assessment should consider EDA hardware, participant sample, experimental context, and the relative importance of keeping vs. removing both valid and invalid data.

There are several strengths of our automated procedure. First, the rules of the automated pipeline are simple and transparent; this makes it easy for future researchers to tune methods to their needs and report their rules so that other researchers can understand how the data were processed. Second, our procedure was developed using a very large EDA dataset (181 hours) acquired in a population from whom it is difficult to obtain high-quality EDA data (children and adolescents with autism), in an ambulatory setting (participants' homes). These conditions are precisely those that require automated quality assessment where manual processing would be time- and cost-prohibitive. Finally, our procedure exhibited excellent agreement with human EDA expert ratings, providing confidence that this procedure will identify invalid portions of data in a manner consistent with human EDA experts, but in a fraction of the time.

Several limitations should be addressed in future work. First, our data were collected only from a population of children and adolescents with autism spectrum disorder. Our automated procedure should be tested with data from different participant populations, other EDA recording devices, electrodes, and recording sites, and from other contexts (home, office, clinic, active vs. sedentary). Second, we did not explicitly detect or analyze SCRs, and our low-pass filters—selected to remove noise—might have diminished SCRs in our data. Indeed, per our study goals, we did not impose rules specifically on SCRs; instead, analyses of SCRs that were in our data were subject to the same four rules as the rest of the EDA data. Third, there is room to improve on our simple rules. For example, it may be possible to account for the slope of temperature data, or the duration during which temperature is out of a desirable range, to determine if the sensor is being worn properly. Moving forward, we will integrate our procedure into a larger EDA processing pipeline that combines the methods developed here with new procedures to distinguish valid SCRs and invalid SCR-like artifacts (such as sensor contact artifacts created when the sensor is touched; e.g., [24, 25]). Lastly, certain EDA artifacts could be corrected using knowledge of EDA phenomena (e.g., [33]) rather than simply excluded, as suggested in prior work (e.g., [34]).

V. Conclusion

Our work fills a well-defined gap in the EDA literature by successfully developing a simple, transparent, flexible, and automated quality assessment procedure for ambulatory EDA data. Our procedure was demonstrated to be effective using a large ambulatory EDA dataset acquired longitudinally from individuals with autism in their homes. Our automated procedure exhibited excellent agreement with quality assessment by multiple independent human EDA experts, but in a fraction of the time required for human coding. To further enhance transparency, rigor, and reproducibility, we provide free and open source MATLAB software to run our procedure at http://www.cbslab.org/EDAQA. We encourage other researchers to replicate and extend our initial efforts. We hope our procedure will enhance the efficiency and transparency of EDA analyses to help advance multiple fields that utilize ambulatory physiological measures, including, but not limited to, clinical studies assessing biomarkers of autism or side effects from cancer treatment, and human factors studies examining the productivity of healthy individuals in their jobs. To paraphrase NIH, scientific rigor and transparency in biomedical research is key to successfully applying knowledge to improve health and well-being [28].

Acknowledgements

The authors would like to acknowledge Caroline Carberry and Amarelle Hamo for their help with data collection, Dr. Stephen Intille and Qu Tang for help with accelerometer analyses, and Dr. Amber Kleckner for helpful feedback in writing this manuscript.

.

References

[1] W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. roth, M. E. Dawson, *et al.*, "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, pp. 1017-34, Aug 2012.

[2] W. Boucsein, *Electrodermal Activity*: Springer Science & Business Media, 2012.

[3] M. van Dooren, J. J. de Vries, and J. H. Janssen, "Emotional sweating across the body: comparing 16 different skin conductance measurement locations," *Physiol Behav*, vol. 106, pp. 298-304, May 15 2012.

[4] R. Kocielnik, N. Sidorova, F. M. Maggi, M. Ouwerkerk, and J. H. Westerink, "Smart technologies for long-term stress monitoring at

work," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, 2013, pp. 53-58.

[5] M. E. Dawson, A. M. Schell, and D. L. Filllion, "The electrodermal system," in *Handbook of psychophysiology*, J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, Eds., ed New York, NY: Cambridge University Press, 2007, pp. 159-181.

[6] F. H. Wilhelm and W. T. Roth, "Ambulatory assessment of clinical anxiety," in *Ambualtory assessment: Computer-assisted psychological and psycho-physioliglca methods in monitoring and field studies*, J. Fahrenberg and M. Myrtek, Eds., ed Gottingen: Hogrefe, 1996, pp. 317-345.

[7] S. F. Morrison, "Differential control of sympathetic outflow," *Am J Physiol Regul Integr Comp Physiol*, vol. 281, pp. R683-98, Sep 2001.

[8] M. M. Bradley and P. J. Lang, "Emotion and Motivation," J. T. Cacioppo, L. G. Tassinary, and G. Berntson, Eds., 3rd ed New York, NY: Cambridge University Press, 2007, pp. 581-607.

[9] M. M. Bradley, "Natural selective attention: orienting and emotion," *Psychophysiology*, vol. 46, pp. 1-11, Jan 2009.

[10] A. Bechara, H. Damasio, A. R. Damasio, and G. P. Lee, "Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making," *J Neurosci*, vol. 19, pp. 5473-81, Jul 01 1999.

[11] S. Geuter, M. Gamer, S. Onat, and C. Buchel, "Parametric trial-by-trial prediction of pain by easily available physiological measures," *Pain*, vol. 155, pp. 994-1001, May 2014.

[12] T. Reinhardt, C. Schmahl, S. Wust, and M. Bohus, "Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the Mannheim Multicomponent Stress Test (MMST)," *Psychiatry Res*, vol. 198, pp. 106-11, Jun 30 2012.

[13] E. B. Prince, E. S. Kim, C. A. Wall, E. Gisin, M. S. Goodwin, E. S. Simmons, *et al.*, "The relationship between autism symptoms and arousal level in toddlers with autism spectrum disorder, as measured by electrodermal activity," *Autism*, Jun 10 2016.

[14] F. H. Wilhelm and W. T. Roth, "Taking the laboratory to the skies: ambulatory assessment of self-report, autonomic, and respiratory responses in flying phobia," *Psychophysiology*, vol. 35, pp. 596-606, Sep 1998.

[15] A. E. Meuret, D. Rosenfield, F. H. Wilhelm, E. Zhou, A. Conrad, T. Ritz, *et al.*, "Do unexpected panic attacks occur spontaneously?," *Biol Psychiatry*, vol. 70, pp. 985-91, Nov 15 2011.

[16] R. G. O'Connell, M. A. Bellgrove, P. M. Dockree, and I. H. Robertson, "Reduced electrodermal response to errors predicts poor sustained attention performance in attention deficit hyperactivity disorder," *Neuroreport*, vol. 15, pp. 2535-8, Nov 15 2004.

[17] M. H. Savard, J. Savard, A. Caplette-Gingras, H. Ivers, and C. Bastien, "Relationship between objectively recorded hot flashes and sleep disturbances among breast cancer patients: investigating hot flash characteristics other than frequency," *Menopause*, vol. 20, pp. 997-1005, Oct 2013.

[18] M. S. Goodwin, W. F. Velicer, and S. S. Intille, "Telemetric monitoring in the behavior sciences," *Behavior research methods*, vol. 40, pp. 328-341, 2008.

[19] F. H. Wilhelm and P. Grossman, "Emotions beyond the laboratory: theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment," *Biol Psychol*, vol. 84, pp. 552-69, Jul 2010.

[20] C. Tronstad, G. K. Johnsen, S. Grimnes, and Ø. G. Martinsen, "A study on electrode gels for skin conductance measurements," *Physiological measurement*, vol. 31, p. 1395, 2010.

[21] H. Storm, A. Fremming, S. Odegaard, O. G. Martinsen, and L. Morkrid, "The development of a software program for analyzing spontaneous and externally elicited skin conductance changes in infants and adults," *Clin Neurophysiol*, vol. 111, pp. 1889-98, Oct 2000.

[22] D. R. Bach, K. J. Friston, and R. J. Dolan, "An improved algorithm for model-based analysis of evoked skin conductance responses," *Biol Psychol*, vol. 94, pp. 490-7, Dec 2013.

[23] J. Blechert, P. Peyk, M. Liedlgruber, and F. H. Wilhelm, "ANSLAB: Integrated multichannel peripheral biosignal processing in psychophysiological science," *Behav Res Methods*, vol. 48, pp. 1528-1545, Dec 2016.

[24] M. Kelsey, A. Dallal, S. Eldeeb, M. Akcakaya, I. Kleckner, C. Gerard, *et al.*, "Dictionary learning and sparse recovery for electrodermal activity analysis," in *SPIE Commercial+ Scientific Sensing and Imaging*, 2016, pp. 98570G-98570G-18.

[25] M. Kelsey, M. Akcakaya, I. R. Kleckner, R. V. Palumbo, L. F. Barrett, K. S. Quigley, *et al.*, "Applications of Sparse Recovery and Dictionary Learning to Enhance Analysis of Ambulatory Electrodermal Activity Data," *Biomedical Signal Processing and Control*, in press.

[26] W. Chen, N. Jaques, S. Taylor, A. Sano, S. Fedor, and R. W. Picard, "Wavelet-based motion artifact removal for electrodermal activity," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 6223-6226.

[27] V. Xia, N. Jaques, S. Taylor, S. Fedor, and R. Picard, "Active learning for electrodermal activity classification," in *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2015, pp. 1-6.

[28] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard, "Automatic identification of artifacts in electrodermal activity data," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 1934-1937.

[29] NIH. (2016, December 12, 2016). *Rigor and Reproducibility*. Available: https://www.nih.gov/research-training/rigor-reproducibility

[30] O. Wilder-Smith and M. S. Goodwin, "Web-Based Toolkit for Multimodal Data Analysis in ASD Research," in *Annual International Meeting for Autism Research*, Salt Lake City, UT, 2015.

[31] J. Cohen, "A coefficient of agreement for nominal scales. Educational and Psychosocial Measurement, 20, 37-46," ed, 1960.

[32] M. E. Rosenberger, W. L. Haskell, F. Albinali, S. Mota, J. Nawyn, and S. Intille, "Estimating activity and sedentary behavior from an accelerometer on the hip or wrist," *Med Sci Sports Exerc*, vol. 45, pp. 964-75, May 2013.

[33] D. R. Bach, G. Flandin, K. J. Friston, and R. J. Dolan, "Modelling event-related skin conductance responses," *Int J Psychophysiol*, vol. 75, pp. 349-56, Mar 2010.

[34] C. Tronstad, O. M. Staal, S. Saelid, and O. G. Martinsen, "Model-based filtering for artifact and noise suppression with state estimation for electrodermal activity measurements in real time," *Conf Proc IEEE Eng Med Biol Soc*, vol. 2015, pp. 2750-3, Aug 2015.