The Current State of EMA and ESM Study Design in Mood Disorders Research: A Comprehensive Summary and Analysis

Meredith L. Wallace1,2, Molly H. Carter2,3 and Satish Iyengar1

1Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.
2Western Psychiatric Institute and Clinic, University of Pittsburgh, Pittsburgh, PA 15213, U.S.A.
3Department of Psychology, Xavier University, Cincinnati, OH 45207, U.S.A.

Abstract. Ecological momentary assessment (EMA) and experience sampling methods (ESM) are becoming increasingly prevalent in mood disorders research due to their potential for capturing underlying dynamic mood processes that cannot be observed through traditional clinical visits. There have also been recent statistical developments that allow for innovative EMA/ESM-related research questions to be answered. However, even the most sophisticated statistical methods cannot glean accurate representations of underlying mood processes when the data are sampled inappropriately. Unfortunately, there are few resources investigators can use to make informed decisions about EMA/ESM study design. Thus, we perform a comprehensive summary of current EMA/ESM study design methods used in mood disorders research, explore the rationale behind study design decisions, and investigate the relationship between compliance and various study design features. Results from these analyses are used to suggest improvements for designing and reporting future EMA/ESM studies.

1 Introduction

Clinical researchers and psychologists base their patient evaluations and diagnoses largely on retrospective self-report of experiences. However, this type of information can be biased by current mood state and day-to-day and even hour-to-hour variability in experiences and symptoms [1–3]. Because mood disorders such as unipolar depression (DEP), bipolar spectrum disorders (BD) and disorders of mood dysregulation such as borderline personality disorder (BPD) are defined by changes in mood state and mood variability [4], recall bias may be a particularly salient problem in this area. As a result, information conventionally collected in clinical settings can fail to capture the continuous, dynamic processes underlying mood disorders, thereby hindering researchers’ abilities to characterize disease processes that may lead to specific and effective treatments. Furthermore, information collected in a clinical setting may not provide a sufficiently detailed understanding of a patient’s course of illness, potentially contributing to the high rate of misdiagnosis across mood disorders [5–9].

Ecological momentary assessment (EMA) [1, 2] and experience sampling methodology (ESM) [10, 11] are data collection methods that allow clinical researchers and
psychologists to capture a patient’s course of illness with greater frequency and sensitivity through the use of hand-held technological devices. When studying disorders of mood and mood dysregulation (henceforth, we will refer to them collectively as “mood disorders”), EMA and ESM are particularly useful for capturing both the mean and instability of mood [12]. EMA accomplishes this through the use of portable technological devices, such as hand-held computers and cellular phones, to capture self-reported measurements as they occur naturally in real time. In this way, EMA allows investigators to control timing precisely and track compliance [13]. EMA can also be used to capture behavioral and physiological data; however, the focus herein is on self-reported EMA. Like EMA, ESM also uses technological devices to capture data in real time; however, ESM typically uses a digital watch or pager to prompt participants to record their experiences in a paper-and-pencil diary [10, 11]. In this manuscript we focus on self-reported EMA and ESM, referring to them collectively as EMA.

As with any type of data collection method, EMA study design is of utmost importance. In general, there is consensus that random sampling should be employed and that the sampling frequency should match the temporal dynamics of the process of interest [11, 14, 15]. However, there is a paucity of formal empirical evidence regarding EMA study design methodology and rationale, particularly as it relates to the capture of underlying mood processes. One important breakthrough study in this area was performed by Ebner-Priemer and Sawitzki [14], who showed that a sampling interval of less than 30 minutes could optimally capture underlying dynamic processes in BPD, while intervals greater than 30 minutes could not. However, few research studies (especially those in clinical populations) are expected to support sampling intervals of less than 30 minutes due to concerns of participant burden [14].

The field of mental health, and mood disorders in particular, could benefit from the development of a more standardized set of EMA study design methods so that future researchers can make more informed decisions (e.g., consider the stringency with which clinical trials are designed and reported). As an initial step towards this end goal we present a comprehensive summary and analysis of study design features in mood disorders research, focusing on three specific aims:

1. Summarize the current state of EMA study design
2. Explore factors that researchers have considered when making EMA study design decisions
3. Investigate the relationships among study design features and compliance

Results from these three aims are used to suggest ways that EMA investigators could improve on the design and reporting of their studies.

2 Methods

To empirically describe and evaluate EMA study design in mood disorders research it is necessary to use the actual studies, rather than manuscripts generated from these studies, as the units of analysis. This presents a challenge because manuscripts and studies are not related on a one-to-one basis. That is, multiple manuscripts may stem
from one study or, alternatively, one manuscript may analyze data from multiple studies. We employed the following strategy to develop a data base with the study as the unit of analysis. First, we set inclusion and exclusion criteria for the types of studies that would be considered. Second, we used literature search engines to locate manuscripts that described studies meeting the inclusion and exclusion criteria. Third, we compared manuscripts to determine which stemmed from the same studies and which merged data from multiple studies. We contacted corresponding authors when questions or ambiguities arose. Fourth, we used the manuscripts and additional resources from authors (e.g., EMA study questionnaires and protocols) to enter the study-level data. We describe these steps in further detail below.

2.1 Study Inclusion and Exclusion Criteria

Study inclusion criteria were: 1) at least one subset of participants was clinically diagnosed with a mood disorder (BPD, DEP, or BD) and 2) an electronic ambulatory device, including but not limited to beepers, hand-held computers, cellular phones, and wrist-watches, was used to capture self-report data at multiple time points. Self-report EMA may be either event-based (participant enters data before and/or after a prespecified event occurs), prompt-based (participant enters data when prompted by a technological device), or a combination of the two; however, the focus herein lies specifically in prompt-based EMA because it relies heavily on a priori decisions regarding sampling frequency. Thus, we excluded studies that did not include some type of prompt-based EMA. We also excluded EMA case studies because of our specific focus on the use of EMA to answer empirically-based research questions.

2.2 Search Strategy

We first aimed to identify all manuscripts arising from appropriate EMA studies by performing literature searches in PubMed, PsycINFO, and ProQuest (Dissertations and Theses) with the following keywords: ambulatory assessment, ecological momentary assessment, experience sampling method, electronic diary, computer-assisted diary, electronic momentary assessment, ecological validity, and hand-held computer. After removing duplicate references, we performed a computerized search of the resulting abstracts to narrow them down to those based on populations with mood disorders. Abstract search terms included depression, depressive disorder, borderline, bipolar, mood disorder, and affective disorder. All remaining abstracts were manually screened to further identify those which stemmed from studies that fit within our inclusion and exclusion criteria. The full text of each of the remaining manuscripts was then used to determine whether the study from which it arose would be considered appropriate for inclusion. Once the final subset of manuscripts was identified, the full text was again examined to determine which manuscripts stemmed from the same EMA study and which combined data from multiple EMA studies.

Figure 1 summarizes the search strategy and manuscript selection process. Table 1 lists the final 27 studies selected through this selection process. Because each of these 27 studies could generate multiple manuscripts, we refer to them in Table 1 by the by the
2.3 Data Collection

The manuscripts referenced in Table 1 and additional author resources (e.g., EMA questionnaires and study protocols) were used to develop a database that included: 1) study design information, such as the sample size, duration of study, number prompts per day, number of questions per prompt, and the type of EMA device used, 2) demographic information for the samples used in each study, such as age and gender, 3) diagnostic groups studied, 4) justification of sampling design, and 5) compliance information, such as missing data and dropout percentages. When collecting the data from studies that also included a randomized trial component, we only included EMA data from the baseline weeks. This was done in an attempt to standardize the studies and also because treatment assignment may impact measures of compliance.

2.4 Data Analysis

We used descriptive statistics to summarize the 27 EMA studies with respect to sampling design and demographic and clinical characteristics of the participants (Aim 1). To explore investigators’ rationale behind EMA study design selection (Aim 2) we first summarized the study design justifications provided in the associated manuscripts. Because only eight of the 27 studies actually provided this explicit justification (see Table 1), we used our observed data to further explore the rationale behind current EMA study design. This involved quantifying observed relationships among study design parameters (the number of days of EMA study, the number of prompts per day, and the number
Table 1. First author and year of the earliest manuscript identified with each study. All other related manuscripts are cited in the “References” column.

<table>
<thead>
<tr>
<th>First Author (Year)</th>
<th>References</th>
<th>First Author (Year)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glaser, J.P. (2008)</td>
<td>[41,42]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Reported rationale for study design

of questions per prompt) using scatter plots and Pearson product moment correlations. Because the distributions of these variables were highly skewed, we used the natural log transformation for all plots and calculations.

To investigate the relationships among study design features and compliance (Aim 3), we first needed to select a single compliance measure. After comparing various possibilities, we chose the percentage of unanswered prompts because it was frequently reported and also easily constructed from other types of reported compliance statistics. We calculated the percentage of unanswered prompts ($P$) as the number of unanswered prompts divided by the total number of possible prompts times 100. Some studies reported only $P_1$, the percentage of unanswered prompts based on a subset of $N_1 < N$ compliant participants with fewer than $X\%$ unanswered prompts. Hence, the percentage of unanswered prompts for the $N_2 = N - N_1$ participants, $P_2$, was unknown except for the fact that $X\% < P_2 \leq 100\%$. Because $N_2$ tended to be very small (e.g., one or two participants), we assumed that $P_2$ followed a uniform distribution and let $P_2 = .5 \times (100 - X)$. $P$ was then calculated as a weighted average of the percentages of unanswered prompts prompts for the $N_1$ compliant and $N_2$ noncompliant participants, that is, $P = \frac{1}{N} (P_1 N_1 + P_2 N_2)$.

After calculating the percentage of unanswered prompts ($P$), we used scatter plots and Pearson product moment correlation coefficients to explore its relationships with various study design features. Specifically, we focused on the number of questions per prompt, the number of prompts per day, and the number of days of the EMA study. Because these study design features need to be carefully balanced to reduce participant burden, we also focused on their interactions: questions per day (questions per prompt × prompts per day), prompts per study (prompts per day × days of study), and questions per study (questions per prompt × prompts per day × days of study). Due to the highly skewed distributions of all variables involved, we used the natural logs of these variables in our plots and calculations.
3 Results

3.1 Aim 1: Summarize the Current State of EMA Research

Study design and demographic characteristics of the of the 27 EMA studies are summarized in Table 2. In general, study participants tended to be female and Caucasian, with a median age of 31. The study design characteristics selected by investigators (e.g., the number of questions per prompts, number of prompts per day, and number of days of EMA) varied widely across the 27 studies, as shown by the minimum and maximum scores in Table 2.

In addition to selecting the number of questions, prompts, and days of EMA, investigators must also choose a method for distributing these prompts throughout the study period. Within each day, the most common method for allocating prompts was random blocking (e.g., divide waking hours into 6 blocks and randomly sample once during each block); this method was used in 10 of the 27 studies (37%). Periodic sampling with random error (e.g., sample every hour plus or minus a randomly drawn number of minutes from a prespecified normal distribution) was the second most common method for allocating EMA prompts, seen in 8 studies (29.6%). Only five (18.5%) studies used random sampling (e.g., randomly select 10 times between the hours of 8:00 am to 10:00 pm) and only four (14.8%) studies used fixed time sampling (e.g., sample every hour on the hour).

The method for distributing prompts throughout the study period must also take into consideration the fact that participants are not able to answer prompts during sleep. Out of 26 studies for which this information was known, 70.1% (n=19) set a fixed daily interval during which prompts could occur (e.g., between the hours of 8:00 am to 10:00 pm for all participants). Other studies set a priori individualized sleep intervals tailored to each participant’s needs (19.2%, n=5) or requested that the participant turn off the device during sleep (7.7%, n=2).

Investigators must also determine which technological device (e.g., hand-held computer, cellular phone, pager) to use to deliver each prompt. A hand-held computer, such as a “personal digital assistant” (PDA) was used in 59.3% of studies (n=16). The next most frequently used technological device (37%, n=10) was a pager or wristwatch along with a paper-and-pencil diary. One study (3.7%) used a cellular phone.

Patients with depression (major depression, minor depression, or dysthymia) were included in eighteen studies (66.7%), patients with bipolar spectrum disorders were included in five studies (18.5%), and patients with borderline personality disorder were included in 9 (33.3%) studies. Other non-affective clinical groups (e.g., schizophrenia, panic disorder) were included in four (14.8%) studies. Healthy controls were used as a comparison group in 16 (59.3%) of the studies.

3.2 Aim 2: Explore Investigators’ Rationale behind EMA Study Design Decisions

We first searched for explicit study design justifications in the manuscripts stemming from each study. Overall, we identified some type of sampling justification in eight of the 27 studies (29.64%). Two of these eight studies discussed rationale for the days on
Table 2. Demographic, clinical, and study design characteristics from 27 studies.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>N</th>
<th>Mean (SD)</th>
<th>Median (Min, Max)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Age</td>
<td>26</td>
<td>31.69 (11.14)</td>
<td>31.04 (10.01, 62.45)</td>
</tr>
<tr>
<td>% Female</td>
<td>24</td>
<td>79.08 (16.54)</td>
<td>78.10 (50, 100)</td>
</tr>
<tr>
<td>% Caucasian</td>
<td>12</td>
<td>65.61 (23.72)</td>
<td>64.70 (33.33, 100)</td>
</tr>
<tr>
<td>% Higher Education</td>
<td>17</td>
<td>56.81 (33.39)</td>
<td>64.50 (0, 100)</td>
</tr>
<tr>
<td>% Married or Cohabitating</td>
<td>15</td>
<td>35.34 (32.00)</td>
<td>26.37 (0, 88.24)</td>
</tr>
<tr>
<td>% Employed (Full- or Part-Time)</td>
<td>12</td>
<td>37.40 (17.77)</td>
<td>43.24 (0, 55.73)</td>
</tr>
<tr>
<td><strong>Study Design</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Size at Study Entry</td>
<td>27</td>
<td>78.67 (43.90)</td>
<td>73 (10, 164)</td>
</tr>
<tr>
<td>Days of EMA/ESM</td>
<td>27</td>
<td>9.33 (9.52)</td>
<td>6.79 (1, 42)</td>
</tr>
<tr>
<td>Prompts per Day</td>
<td>27</td>
<td>9.16 (9.63)</td>
<td>8 (1, 54)</td>
</tr>
<tr>
<td>Questions per Prompt</td>
<td>26</td>
<td>25.58 (16.91)</td>
<td>23 (1, 75)</td>
</tr>
<tr>
<td>Percentage of Missed Prompts</td>
<td>22</td>
<td>16.79 (10.40)</td>
<td>12.97 (3, 41.9)</td>
</tr>
</tbody>
</table>

*Among 24 studies with an average participant age > 18

which EMA sampling occurred, citing that “the weekend was chosen because it is the time when adolescents have the greatest amount of free time and control over activities and companions” [21] and “the same weekdays (Tuesday-Thursday) were used to have a homogenous sample of days” [29].

The remaining six of the eight studies provided justification for the timing and/or frequency of prompts within each day. In two different studies, Stetler et al. [56, 57] sampled cortisol levels at the same time as the self-reported EMA. Specific sampling time intervals were chosen because they were previously found to “...adequately capture the diurnal pattern of cortisol secretion without placing undue burden on the participants” [56] and because they were able to “…capture the early morning peak that is part of the diurnal pattern of cortisol secretion” [57]. Ebner-Priemer and Sawitski [14] emphasized that “the temporal dynamics of emotional-cognitive processes are largely unknown”, and thus, their study employed multiple sampling frequencies to investigate this question. Depp et al. [34] cited the “need to balance between ‘coverage’ of affective experiences and subject burden.” Doyle [35] simultaneously employed three different types of recording procedures “to capture mood ratings in close proximity to the behaviors and events of interest...”. Links [44] stated that “random times were used to approximate the daily range of a participant’s affective intensity within the context and flow of the participant’s daily experience.”

Because there were only eight studies for which an explicit sampling design rationale was found, we also explored the observed relationships among the number of days of EMA, the number of prompts per day, and the number of questions per prompt. Because the study design variables were highly skewed, Figure 2 displays the associations among the log-transformed study design variables (henceforth, the reader may assume that all variables discussed are logged transformed). There was a strong negative association between the number of questions per prompt and the number of prompts per day ($r = -.45, p = .02$). Similarly, there was a strong negative association between the number of prompts per day and the number of days of EMA ($r = -.41, p = .03$). There
was a strong positive association between the number of questions per prompt and the number of days of EMA ($r = .48, p = .01$). The positive association may reflect the fact that more days of EMA leads to fewer prompts per day, which may in turn lead to more questions asked at each prompt. There was no significant association between the number of questions per prompt and the total number of prompts over the entire EMA study.

3.3 Aim 3: Investigate Relationships among Study Design Features and Compliance.

Figure 3 illustrates the relationships among the study design features and the percentage of unanswered prompts. The number of questions per prompt and the number of prompts per day were not significantly associated with the percentage of unanswered prompts. After removing the high but valid outlier in the number of prompts per day (estimated 54 prompts per day [36]), there was a strong positive relationship between the percentage of unanswered prompts and the number of prompts per day ($r = .44, p=0.05$); however, the removal of this outlier only highlights the potential leverage of the low outliers (1 prompt per day [56, 57]). There were strong and borderline-significant positive relationships between the percentage of unanswered prompts and both the days of EMA and questions per day (questions × prompts). There were strong and significant positive relationships between the percentage of unanswered prompts and both the total number of prompts (prompts × days) and the total number of questions (questions × prompts × days).
Fig. 3. Scatter plots with least-squares regression lines displaying relationships between the log-transformed study design features and the log percentage of unanswered prompts. Pearson product moment correlation coefficients (r) and associated p-values are also displayed.

4 Discussion

The overall goal of this manuscript is to provide researchers with a comprehensive summary and analysis of current EMA study design methods used in mood disorders research. To attain this goal, we summarized the current state of EMA study design used in mood disorders research, explored the rationale behind the selection of EMA study design features, and investigated the impact of study design on participant compliance.

The results of our comprehensive summary highlight the wide variety of EMA study design methods that are currently used in mood disorders research. This is particularly true regarding to the number of questions per prompt, the number of prompts per day, and the number of days of EMA. To a large degree, this variability may be explained by the fact that each study has its own unique goals, participants, and restrictions; thus, each study’s design must be tailored to meet its specific needs. However, insight into this process was only provided to the reader in 8 of the 27 studies we investigated. This lack of detail may pose challenges for newer investigators who may want to enter into EMA research but do not have the experience to make these decisions on their own. It may also lead more experienced EMA researchers to use only one familiar study design, rather than tailoring each study design to match the temporal dynamics of the underlying process of interest.

The lack of explicit detail regarding study design rationale makes it difficult to explore which considerations are most important for EMA investigators. However, the observed negative associations between the number of prompts per day and both the
days of EMA and the questions per prompt suggests that investigators do indeed consider the need to balance these study design features, presumably to reduce participant burden.

Our investigation of the relationships among study design features and participant compliance showed that the number of study days may have a bigger impact on compliance than either the number of prompts per day or the number of questions per prompt. Not surprisingly, the strongest observed relationship showed a positive association between the total number of questions asked during the study (questions × prompts × days) and the percentage of unanswered prompts. Although these relationships may not be unexpected, their quantification is an important first step towards developing a set of study design guidelines for future EMA research.

Although care was taken to avoid potential biases and when developing the data base, analyzing the data, and interpreting the results, there are limitations that must be considered. One such limitation stems from the fact that our primary unit of analysis was the study. As such, features of each study often had to be identified through manuscripts and by discussion with the actual study investigators. When investigators could not be reached, the full scope of the original EMA study was not always evident; this could lead to errors in the data base due to a lack of full study descriptions in these manuscripts. However, the fact that these study design features were not immediately evident from the manuscripts highlights the need for a more standardized approach to developing and reporting EMA studies.

4.1 Future Directions

When designing EMA studies, two critical considerations are to obtain quality EMA data at the frequency necessary for modeling underlying dynamic processes and to use a sampling scheme that will not result in undue participant burden [14]. These considerations are often at odds with one another, and thus, pose challenges for EMA researchers. Furthermore, the “underlying temporal dynamics” of mood disorders are still largely unknown [14], making it difficult to select the appropriate sampling frequency even without participant constraints.

To overcome these challenges, it will be important to address three areas of EMA research: 1) monitor and standardize current sampling methods 2) evaluate whether current EMA sampling methods work actually work as intended (i.e., capture true underlying processes), and 3) develop new EMA sampling methods that can balance the need to effectively capture these processes while considering participant burden. The research presented herein is aimed at addressing the first area of research. We are currently working to address areas two and three so that future EMA investigators can have the tools they need to answer the critical EMA-related questions, both in the field of mood disorders specifically and in the broader fields of mental and physical health research.

References


