AIDS Behav (2015) 19:330–340
DOI 10.1007/s10461-014-0923-8

ORIGINAL PAPER

Reliability and Validity of Daily Self-Monitoring by Smartphone Application for Health-Related Quality-of-Life, Antiretroviral Adherence, Substance Use, and Sexual Behaviors Among People Living with HIV

Dallas Swendeman · W. Scott Comulada · Nithya Ramanathan · Maya Lazar · Deborah Estrin

Published online: 21 October 2014

Abstract This paper examines inter-method reliability and validity of daily self-reports by smartphone application compared to 14-day recall web-surveys repeated over 6 weeks with people living with HIV (PLH). A participatory sensing framework guided participant-centered design prioritizing external validity of methods for potential applications in both research and self-management interventions. Inter-method reliability correlations were consistent with prior research for physical and mental health quality-of-life ($r = 0.26$–0.61), antiretroviral adherence ($r = 0.70$–0.73), and substance use ($r = 0.65$–0.92) but not for detailed sexual encounter surveys ($r = 0.15$–0.61). Concordant and discordant pairwise comparisons show potential trends in reporting biases, for example, lower recall reports of unprotected sex or alcohol use, and rounding up errors for frequent events. Event-based reporting likely compensated for modest response rates to daily time-based prompts, particularly for sexual and drug use behaviors that may not occur daily. Recommendations are discussed for future continuous assessment designs and analyses.

Keywords Self-monitoring · mHealth · Reliability · Validity · HIV/AIDS

Introduction

Mobile phones are increasingly being advocated for innovation in psychosocial and behavioral health research and interventions as part of a broader "mHealth" agenda [1, 2]. While there is a noted lack of an evidence-base for mHealth methods [1, 2], mHealth is moving forward rapidly in research and commercial applications. Mobile phones are enabling the rapid and inexpensive deployment of previously un-scalable methods for daily and in-the-moment assessments of states, behaviors, and experiences, such as ecological momentary assessment (EMA) and daily diaries [3–6]. These methods enable examination of daily variations in events, behaviors, states, and their co-variation, examination of individual variation including in treatment responses, more refined causal inferences [3–6] and potentially delivery of personalized and in-the-moment interventions [7].

One of the major challenges identified in advancing mHealth methods and evidence is the establishment of reliability and validity of measures [2], with few studies reporting on reliability and validity of daily self-reports. This study examines the reliability and validity of mobile self-reports by PLH in a pragmatic and participatory [8] pilot study of a mHealth self-monitoring platform tailored

Electronic supplementary material The online version of this article (doi:10.1007/s10461-014-0923-8) contains supplementary material, which is available to authorized users.

D. Swendeman · W. S. Comulada · M. Lazar
Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, USA

D. Swendeman (✉)
Center for HIV Identification, Prevention, & Treatment Services (CHIPTS), UCLA, 10920 Wilshire Blvd. Suite. 350, Los Angeles, California 90024, USA
e-mail: dswendeman@mednet.ucla.edu

N. Ramanathan
Department of Computer Sciences, University of California, Los Angeles, USA

D. Estrin
Department of Computer Sciences, Cornell Tech, New York, NY, USA

for HIV-related domains, with an ultimate aim of informing future assessment design and analysis methods for applications in both research and self-management interventions. The study aimed to compare daily self-reporting by smartphone to less frequent and possibly more sustainable self-monitoring via 2-week retrospective recall web-surveys. Inter-method reliability (i.e., consistency of reports) is assessed by examining agreement in terms of correlations and concordant/discordant pairwise comparisons. Validity (i.e., accuracy of reports) is also examined through comparisons of concordant/discordant pairs of reports and visual plots showing trends in over- or under-reporting. The paper concludes with recommendations for design of future mHealth and web-based self-monitoring tools and analytic considerations, including assessment of reliability and validity.

Daily assessments (i.e., EMA, diaries) are broadly considered to be the gold standard when compared to more traditional retrospective recall assessments because daily reports are less subject to recall bias and have greater ecological validity [3–7]. Recall bias and the reliability of self-reports vary based on the level of detail queried, frequencies of events or states being reported, social desirability, administration format, anchoring techniques, recall periods, and protocol compliance or completion rates [3–7]. Compared to retrospective reports, daily assessments typically have fewer questions but much higher frequency burden. This burden may decrease participation rates (i.e., compliance or completion rates) [3–7], and in turn, impact reliability, validity, biases, and inferences due to data that may not be missing at random [9]. Relatively few studies using daily assessment methods address reliability and validity. This paper aims to help fill this gap in the evidence-base on mobile health and daily assessment methods [2].

Although there are large research literatures using daily assessments for physical and mental health symptoms [10–12], including among PLH and gay/bisexual men [13, 14], relatively little has examined reliability or validity. Inter-method reliability correlations for affect measures, for example, range from r = 0.30 to 0.90 depending on specific measure and statistical methods used [15, 16]. Mobile phone reporting has also been used recently for real-time antiretroviral therapy (ART) adherence measurement, which finds low reliability of self-reports [17–19]. Frequent and undesirable substance use behaviors (e.g., smoking or drinking by participants in cessation programs) typically have higher recall reports compared to daily reports due to rounding up errors, with discrepancies increasing with greater frequency or regularity of use [20–22]. A relatively large literature examines daily-recall comparisons for self-reports of sexual behaviors finding reliability correlations ranging from r = 0.87 to 0.97 [23–31] but with recalls generally lower than daily reports, particularly for risky behaviors such as unprotected anal intercourse [23, 30, 31]. The results of these studies and the proponents of daily assessment methods broadly agree that EMA and daily diaries improve accuracy of self-report, particularly for common and frequent experiences [3–7].

Formative Research & Preliminary Results

Formative focus groups with PLH ($n = 29$) recruited from the primary study site informed the design and anticipated challenges with using mobile phone and web surveys [32]. Participants suggested that compliance and reliable reporting would be challenging for some PLH in some domains, particularly sexual behaviors and substance use due to burden and privacy concerns but mitigated by financial incentives. PLH also expressed interest and motivation to self-monitor for self-discovery, self-management, and sharing information with service providers [32].

Prior reports of the current study's mixed-methods data (e.g., brief open-ended interviews at follow ups) demonstrate feasibility, acceptability, and participant perceptions of the efficacy of mobile self-monitoring for self-awareness and self-management [33]. About 50 % of participants reported increased awareness and about 25 % reported behavioral changes or therapeutic benefits, across domains, in response to mobile self-monitoring, with rates tending to be lower for less frequent behaviors [33].

Methods

This study aimed to examine daily self-monitoring by mobile phone application guided by a participatory sensing perspective used in mobile phone sensing projects [8], in which an application for a topic is developed, participation is invited from a community, and observations are made on how people use the tools given varying ability and motivations. The utility of the data and user experiences are assessed for both researchers and participants. Similar to pragmatic designs in implementation research [34], participatory sensing's emphasis on naturalistic use prioritizes external validity and generalizability of the tool use across diverse user preferences, participation options, and motivations for participation. In the current study, participatory sensing guided decisions around incentives for phone surveys, event-based and time-based reporting instructions, and low burden non-response options.

Recruitment, Eligibility, Screening, & Randomization

Fliers were posted at two AIDS service organizations in Los Angeles. Per UCLA institutional review board

requirements, the study flier listed eligibility criteria, study purpose, and contact phone number. Interested agency clients called the study phone number, were screened for eligibility, and if eligible, were invited to an in-person appointment at the site to complete informed consent, baseline interview, and review study instructions.

Eligibility criteria for the study included self-reported: HIV + status; current alcohol, tobacco, or other drug (ATOD) use and sexual activity (at least once/week); daily mobile phone and internet usage; written and verbal English fluency; daily medication use; current client at recruitment site; 18 years of age or older; and having stable housing and source of income. The latter two criteria were informed by the focus groups, which suggested that people without stable housing and income would have high incentives to sell the mobile phones provided by the study. However, the thresholds for stable income and housing were low, and included public assistance, disability payments, public housing and staying with friends or relatives.

Over a nine-month recruitment period the study coordinator received 126 calls, screened 118 individuals, and found 53 were eligible, of which 50 participants were consented and enrolled. Eligible participants were randomized into one of three groups; two mobile phone groups ($n = 34$) and a web-survey only comparison group ($n = 16$). Randomization lists were balanced across self-reported ethnic (African-American, Latino, Caucasian/Other, Asian/Other) and gender categories (i.e., block randomized). The comparison group (not examined in this analysis) was for a preliminary aim to examine behavioral changes in response to smartphone self-monitoring, which did not show effects in preliminary statistical analyses. The two mobile phone groups also had minor variation in consent forms, with two paragraphs framing the study as either focused on developing a new research tool ($n = 14$) or as a behavior change tool ($n = 20$) to preliminarily examine potential impact on participation. Preliminary analyses of both quantitative and qualitative data indicated that there was insufficient emphasis after consent for participants to recall the framing or to observe differences in statistical trends, although the study was not powered statistically to detect significant changes.

Procedures

At the first in-person meeting, all participants completed a baseline retrospective computer-assisted self-interview (CASI) on Survey Monkey. The research assistant (RA) provided brief instructions and was present to answer questions. Participants signed equipment sign-out forms and received study assigned mobile phones (a first generation Android G1 smartphone, $50 street value). The RA trained participants in phone use (including security lock) and the mobile application (*Ohmage,* www.ohmage.org), which included review of phone surveys and written instructions with screen-shots. The RA also assisted participants in customizing time-based smartphone survey prompts (alarms).

Study activities consisted of 6 weeks of daily smartphone self-monitoring divided into three two-week periods. All participants were scheduled to complete the following:

- Web surveys at baseline, and end of Weeks 2, 4, and 6. Email reminders contained personalized survey links. The RA called participants when surveys were not completed within 3 days of the due date. Per IRB requirements, any question could be skipped (i.e., refused) and most also provided "don't know" response options.
- Mobile phone surveys once daily on ATOD use, sexual behaviors, and medication adherence, and four times-per-day on physical and mental health-related quality-of-life (HRQOL). Participants were instructed to complete phone surveys when prompted by the application alarm (time-based reporting) and whenever relevant experiences occurred (event-based reporting).
- Qualitative user-experience interviews (not used in the current analysis) by phone at end of Weeks 2 and 4, and in-person at the final 6 week meeting [33].

Participants were compensated up to $170: $25 for in-person meetings at baseline and 6 weeks; $10 for each follow-up web survey; $10 for each phone interview; and $5 for completing 25 % of phone surveys, $15 for 50 %, $20 for 75 %, and $30 for 100 %. Incentives were provided for seven phone assessments per day, with the recommended schedule being four HRQOL surveys throughout the day and once daily substance use, sexual behavior, and medication adherence surveys (typically end of day).

2-week Recall Web-Surveys

Demographics assessed at baseline included age, gender, sexual orientation, race/ethnicity, education, and number of hours per week working or volunteering.

Health-Related Quality of Life (HRQOL). The Centers for Disease Control and Prevention's brief HRQOL measure was used to assess physical and mental health symptoms, which has good to excellent retest reliability at 0.75 or higher [10]. The current analyses examines the Healthy Days Symptoms Module, which consists of 5 questions (Cronbach's alpha = 0.72 based on baseline data) on number of days experienced: 1) depression ("sad, blue, depressed"), 2) anxiety ("worry, tense, or anxious"), 3) physical symptoms of energy level ("very healthy and full

of energy"), 4) fatigue ("not get enough rest or sleep"), and 5) activity limitations ("usual activities were hard to do") due to pain or poor physical or mental health. Questions were modified to assess the past two weeks.

Antiretroviral medication adherence was assessed using the AIDS Clinical Trial Group (ACTG) adherence questionnaire [35], which was modified to also assess the prior two-weeks. We examined 3-day and 14-day recall responses for ART adherence.

Substance use was assessed using measures used in prior studies with PLH [36]. Questions assessed numbers of days of use over the prior 2 weeks for alcohol, tobacco, marijuana, cocaine, crack, methamphetamine/stimulants, hallucinogens, and heroine/opiates.

Sexual Behaviors and HIV risk were assessed using a slightly modified version of the NIMH Multisite Prevention Trial Protocol assessment [37], adapted to assess the prior two-weeks. Questions used in this analysis assess total number of sex partners and partner-level reports for up to five recent sex partners on numbers of sex acts, unprotected sex acts, and unprotected sex acts with HIV- or unknown status partners.

Mobile Phone Surveys

Mobile phone surveys were adapted from web-survey questions to assess behaviors and states on a daily basis. The surveys were organized in the app into the four categories outlined below. Questions could be skipped by pressing a "skip" or "next" button on each page of the application.

Physical & Mental Health Symptoms (HRQOL; 5 items; prompted 4x/day). Five HRQOL items were adapted from the web-based surveys for EMA based on expected variability throughout a day. Each item was rated on a 0–3 scale (Cronbach's alpha = 0.73, based on the first observation for each person). Two classification thresholds were examined for comparisons with recall days' reports: 1) 'Not at all' versus 'A little', 'Somewhat', or 'Extremely'; and 2) 'Not at all' or 'A little' versus 'Somewhat' or 'Extremely,' with the latter showing higher agreement with recalls and so used in this analysis.

Medication adherence (8 items; prompted 1x/day). Questions included whether a medication was missed or taken. Only reports on ART are used in this analysis due to variability in other medication use reports. Detailed questions for other analysis aims beyond the scope of this paper included if took medication on time, and reason for late or missed doses.

Alcohol, Tobacco, other Drugs (ATOD; 12 items with branching options; prompted 1x/day). A check all that apply stem question queried whether alcohol, tobacco, marijuana, cocaine, methamphetamine, or "other" substances were used "since last report" (i.e., prior day ideally). This time framing was used to anticipate missed daily reports and event-based reporting trends.

Sexual Encounters (17 items with branching options; prompted 1x/day). Sexual encounter information assessed since last report included: partner type (e.g., one-time or regular), gender, HIV and sexually transmitted infection status, and nickname for repeat reports; time since encounter ended; anal, vaginal or oral intercourse; active or receptive partner; condom use; safe sex discussions; and ATOD use during the encounter. The final item instructed participants to repeat the survey for each sexual encounter. For comparison with the web survey, we created indicator variables as a proxy for sexual behaviors on a day (1) or not (0).

Data Analysis

Correlation analyses focus on comparisons of daily mobile and two-week (14-day) web-survey recall reports, with up to three recall periods per participant. Agreement correlations are calculated using methods similar to Carney et al. [38] and Shrier et al. [26]. Strengths of association are estimated by Spearman correlation coefficients (r), which is a nonparametric version of the Pearson correlation coefficient and appropriate for measures in this data with skewed distributions. To account for high potential correlations due to zero estimates (e.g., non-smokers consistently reporting zero smoking days), correlation coefficients are also calculated and presented that exclude non-users of specific substances and those sexually abstinent during a recall period.

Concordance and Discordance. Correlations indicate inter-method reliability (i.e., consistency between daily and recall reports), but consistent over- or under-reporting can still exhibit high correlations. Validity, to complement reliability, refers to the precision or accuracy of the measurement and is indicated by daily and recall comparisons that do not show consistent over or under reporting. In this study we assess agreement in terms of *one-to-one correspondence* between recall and diary reports by examining differences in daily and recall report means, concordant and discordant pairs, and visual plots showing the combination of correlations, means, and concordance.

Observations nested within subjects. Since the unit of analysis is 14-day recall periods over six weeks, with up to three periods nested within individuals, random-effect models are used to calculate p values for mean differences that account for possible intra-cluster correlation between recall periods j nested within participant i. The model is expressed as: $\eta_{ij} = \beta_0 + \beta_1 \text{ x } \text{Daily}_{ij} + \lambda_i$, where η_{ij} is either a logistic link function for proportions or a logarithmic link function for counts, β_0 and β_1 fixed effects, λi

is a participant-level random effect, and $Daily_{ij}$ is an indicator variable for whether the number of events is based on daily reports ($Daily_{ij} = 1$) or recall report ($Daily_{ij} = 0$). A significant mean difference is indicated by a significant β_1 effect. Logistic regression with random-effects is fit to proportions, except for tobacco use due to convergence problems. Because counts contain a high fraction of zeros, negative binomial random-effect regression is used to account for over-dispersion where variances of counts are much larger than mean counts. Poisson random-effect regression is fit when convergence problems are encountered. Models without random-effects likely underestimate standard errors, so significant results are interpreted with caution for those models.

Analyses are carried out in SAS software version 9.3. [39]. Spearman and Pearson correlation coefficients are calculated in the PROC CORR procedure, random-effect logistic models are fit in the NLMIXED procedure, logistic models without random effects are fit in the LOGISTIC procedure, and both random-effect negative binomial and Poisson models are fit in the GLIMMIX procedure. Plots are produced using R [40].

Missing Data

Analyses examine both proportions of days and absolute *counts* of days reporting events over 14-day recall periods. Sexual behaviors analyses only examine counts because the recall and daily surveys are event-level measures. Analysis of counts of days reported assumes event-based reporting with a general, but uncertain, assumption that missing data is at random and ignorable. Analyzing comparisons of proportions of days reported assumes time-based reporting with missing data assumed to be ignorable (i.e., at random) by adjusting for differing numbers of days in the denominators; results are the same as if mean imputation methods were used for missing day reports. Notably, if daily reporting compliance is low then the ignorable missing data assumption is less reliable. For example, a participant may recall 7 days of alcohol use (50 %) and similarly report alcohol consumption on 5 of 10 days of diary reports (50 %) but 36 % if the denominator is all 14 days in the period. If only one day of diary reporting is completed, for example, the proportion can only take on values of either 0 or 100 %. Therefore, to provide further adjustment for an ignorable missing data assumption, analyses exclude 14-day periods with less than 7 days of daily reports.

Non-ignorable missing data patterns are difficult to adjust for, but two steps are taken to detect their presence. First, daily and recall survey comparisons based on both proportions and counts are examined to reflect both event-based and time-based reporting assumptions, respectively. For example, if surveys are only completed on use days (i.e., event-based reporting only), a participant may report 7 days of substance use on both daily and recall surveys. The proportion difference (time-based assumption) is 100 % mobile days - 50 % recall days = 50 % but the absolute count difference (event-based assumption) is 7−7 = 0. Therefore, plots are also presented of correlations between daily-recall count discrepancies and the number of days missing daily reports using the full data set (i.e., not excluding low daily survey compliance periods). A negative relationship between the daily-recall difference and missing days is indicative of daily under-reporting when use is occurring. A large number of missing days may correspond to a heavy substance user who under-reports daily but has higher recall reports. A positive relationship between daily-recall differences and missing days also suggest a missing data mechanism but it is more difficult to posit a particular cause for such a pattern.

Results

Participants' ($n = 34$) ages averaged in the mid-forties (range = 23–64 years old) and most reported taking ART (79 %). Participants were African American (44 %), White (29 %), Latino (12 %), mixed-ethnicity (12 %), and Native-American (3 %). Most were male (77 %); gay (64 %) or bisexual (18 %); reported graduating high school or obtaining a GED certificate (61 %) or a college degree (21 %). Half (50 %) were currently employed and the remainder had income from disability or similar programs.

Recall and Daily Survey Completion Rates

About 70 % (24/34) completed all three follow-up web-surveys, 15 % (5/34) completed only two follow-up surveys, 3 % (1/34) completed only one follow-up survey, and 12 % (4/34) did not complete any and were excluded from analyses. Missing data within surveys from skipped and "don't know" responses was also relatively high compared to the team's experience with in-person assessments in prior studies [37], likely due to ease of non-response on web-surveys. Daily phone survey participation rates were 50 % (17/34) reporting for 6 weeks, 24 % (8/34) for 4 weeks, 15 % (5/34) for 2 weeks, and 12 % (4/34) for less than 1 week. The median number of days reported within each 14-day recall period was 10 days for HRQOL, 7 days for medication use, 7–8 days for ATOD use, and 8 days for sexual behaviors. Excluding 14-day periods with less than 7 days of daily reporting resulted in data set for analyses consisting of $n = 61$ recall periods (range $n = 34$–53 due to missing data) across N = 26 participants (see Tables 1 and 2).

Table 1 Summary of agreement between (N) pairs of daily (D) and recall (R) reports for analyses of *absolute counts of days* reported over 14-day recall periods

Measure	N	Mean count		*t* test[a]	Concordant pairs[b]		Discordant pairs		Correlation[c]	
		D	R	t_{df}	0	Non-0	D > R	D < R	0	Non-0
Health-related QoL										
Fatigue (lack rest)	43	5.5	4.6	1.88_{62}	3	4	22	14	0.51	
Healthy (energy)	39	7.2	5.7	2.49^{*}_{55}	3	1	24	11	0.57	
Depression	48	3.5	5.2	-4.05^{**}_{73}	5	4	12	27	0.59	
Anxiety	48	5.1	5.8	-1.57_{70}	4	6	16	22	0.59	
Usual Acts Hard (due poor health)	44	5.3	2.5	6.29^{**}_{63}	3	4	29	8	0.26	
Usual Acts Hard (due to pain)	53	5.0	3.4	3.92^{**}_{81}	5	2	31	15	0.23	
ART Non-Adherence										
14 days	34	1.3	2.2	-1.30^{d}_{49}	18	2	6	8	0.69	0.40
3 days	36	0.3	0.4	-0.99^{d}_{56}	28	2	2	4	0.73	−0.03
ATOD Use										
Alcohol	48	2.8	3.2	-0.31^{d}_{72}	13	5	16	14	0.65	0.43
Tobacco	44	4.8	6.0	$-2.39^{*,d}_{63}$	22	2	3	17	0.92	0.48
Marijuana	46	2.8	3.4	-1.60_{68}	24	0	9	13	0.84	0.28
Cocaine	47	0.7	0.9	-1.37_{69}	34	2	5	6	0.73	0.40
Meth	47	0.6	0.8	-0.85_{69}	35	1	5	6	0.80	0.42
Sexual behavior										
Partners	41	0.4	1.3	$-3.76^{**,d}_{61}$	17	6	2	16	0.59	0.40
Acts	45	2.4	2.9	-1.49_{67}	19	2	5	19	0.61	0.36
Unprotected acts	47	1.8	1.3	2.15^{*}_{70}	27	1	7	12	0.54	0.18
Unprotected acts w/HIV-/unknown	46	1.0	0.8	0.87_{68}	33	0	4	9	0.15	−0.44

* $p < 0.05$; ** $p < 0.01$

[a] *t* test statistics (degrees of freedom; t_{df}) for comparison between D and R mean counts

[b] Concordant pairs for zero reports (0) and non-zero (Non-0) reports (i.e., exact match)

[c] Spearman correlation coefficients between D and R reports with zero reports included (0) and excluded (Non-0)

[d] Negative-binomial model; Poisson model fit to other measures

Demographic characteristics were compared between participants in the analysis sample ($n = 26$) and those excluded ($n = 8$). Excluded participants were more likely to be younger (mean = 38.9 vs. 47.3; t = −2.23, df = 32, $p = 0.03$) and showed a trend towards being African American (75 vs. 35 %; Fisher's Exact test, $p = 0.10$).

Agreement Between EMA and Recall 14-day Periods

Tables 1 and 2 show daily-recall inter-method comparison results for counts of days and proportions (percentages) of days, respectively, including concordant and discordant pairs, and Spearman correlation coefficients. Results are visualized for selected representative variables in Fig. 1 (with jitter added to differentiate overlapping points). Perfect agreement is indicated by points that fall on a 45-degree line, with points clustered at the low end of the diagonal indicating zero reports (i.e., abstinence) during a period. Plots with points that are clustered consistently below or above a diagonal indicate under- or over-reporting trends, respectively (i.e., reliable but less valid or accurate). As shown in the tables and plots, reliability correlations are lower when excluding zero-reports (i.e., abstainers).

HRQOL

HRQOL daily reports were higher compared to recall for activity limitation, fatigue, and energy, as reflected in comparisons of means and discordant pairs (see Tables 1 and 2). By contrast, anxiety symptoms had balanced reports, and depressive symptoms had lower reports for daily compared to recall. Figure 1 shows plots for depressive symptoms to illustrate (i.e., points below diagonal). Consistency of results across mean count and proportion analyses (Table 1 compared to Table 2) is a result of high daily HRQOL survey participation rates. Correlations are modest (r = 0.51–0.61) except for activity limitations (r = 0.23–0.31), with the latter likely due to low daily reporting (averaging 3 days) and reflecting lack of reporting non-events.

Table 2 Summary of agreement between (N) pairs of daily (D) and recall (R) reports for analyses of *proportion of days* reported over 14-day recall periods

Measure	N	Mean Percent D	Mean Percent R	*t* test[a] t_{df}	Concordant pairs[b] 0	Concordant pairs[b] Non-0	Discordant pairs D > R	Discordant pairs D < R	Correlation[c] 0	Correlation[c] Non-0
Health-related QoL										
Fatigue (lack rest)	43	49.5	33.2	6.38^{**}_{22}	3	3	26	11	0.51	
Healthy (energy)	39	63.6	41.0	8.05^{**}_{21}	3	0	30	6	0.61	
Depression	48	31.6	37.3	-2.32^{*}_{21}	5	1	16	26	0.60	
Anxiety	48	45.8	42.2	1.42_{24}	4	3	22	19	0.60	
Usual Acts Hard (due poor health)	44	47.6	18.3	10.62^{**}_{23}	3	0	34	7	0.29	
Usual Acts Hard (due to pain)	53	45.8	24.7	8.81^{**}_{23}	5	0	35	13	0.31	
ART Non-Adherence										
14 days	34	13.6	15.7	-0.85_{17}	18	0	8	8	0.70	0.43
3 days	36	10.2	10.4	-0.16_{14}	28	0	5	3	0.73	0.06
ATOD Use										
Alcohol	48	27.9	22.9	2.44^{*}_{22}	13	0	22	13	0.65	0.44
Tobacco	44	43.8	42.7	1.37^{d}_{1}	22	13	5	4	0.94	0.53
Marijuana	46	26.6	24.2	1.64_{22}	24	1	10	11	0.82	0.14
Cocaine	47	6.6	6.8	-0.19_{23}	34	1	7	5	0.75	0.45
Meth	47	6.0	5.7	0.18_{23}	35	0	7	5	0.79	0.24

* $p < 0.05$; ** $p < 0.01$

[a] *t* test statistics (degrees of freedom; t_{df}) for comparison between D and R mean proportions

[b] Concordant pairs for zero reports (0) and non-zero (Non-0) reports (i.e., exact match)

[c] Spearman correlation coefficients between D and R reports with zero reports included (0) and excluded (Non-0)

[d] Chi square statistic from logistic regression without random effects

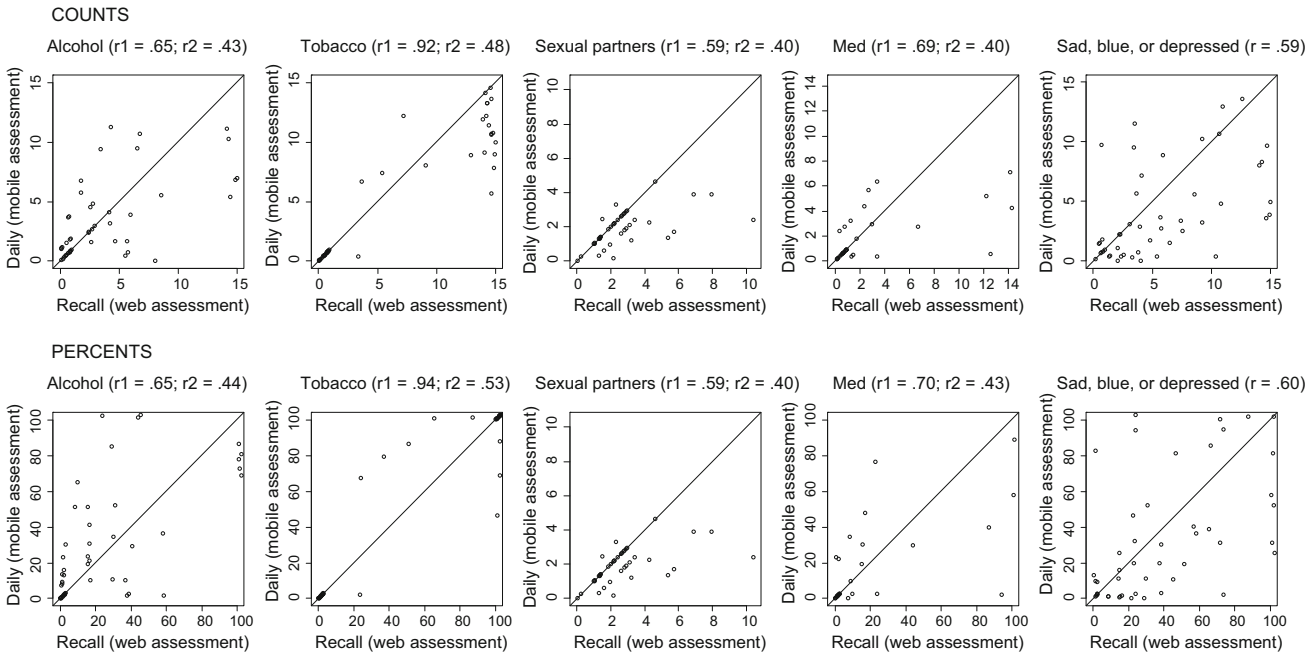

Fig. 1 Plots of exemplar domains showing daily and recall pairwise comparisons for counts and proportions of days reported, and corresponding Spearman's rho correlations shown for 14-day periods including zero reports (r1) and excluding zero reports (r2)

Antiretroviral Medication Adherence

Mean count and proportions of days missing at least one HIV medication were similar for recall compared to daily report, and for 14- or 3-day recall periods (Tables 1 and 2). Correlations were relatively high when including 100 % adherent participants (i.e., zero missed dose reports) at about r = 0.70. Discordant pairs comparisons show fairly balanced over- and under-reporting (also see Fig. 1).

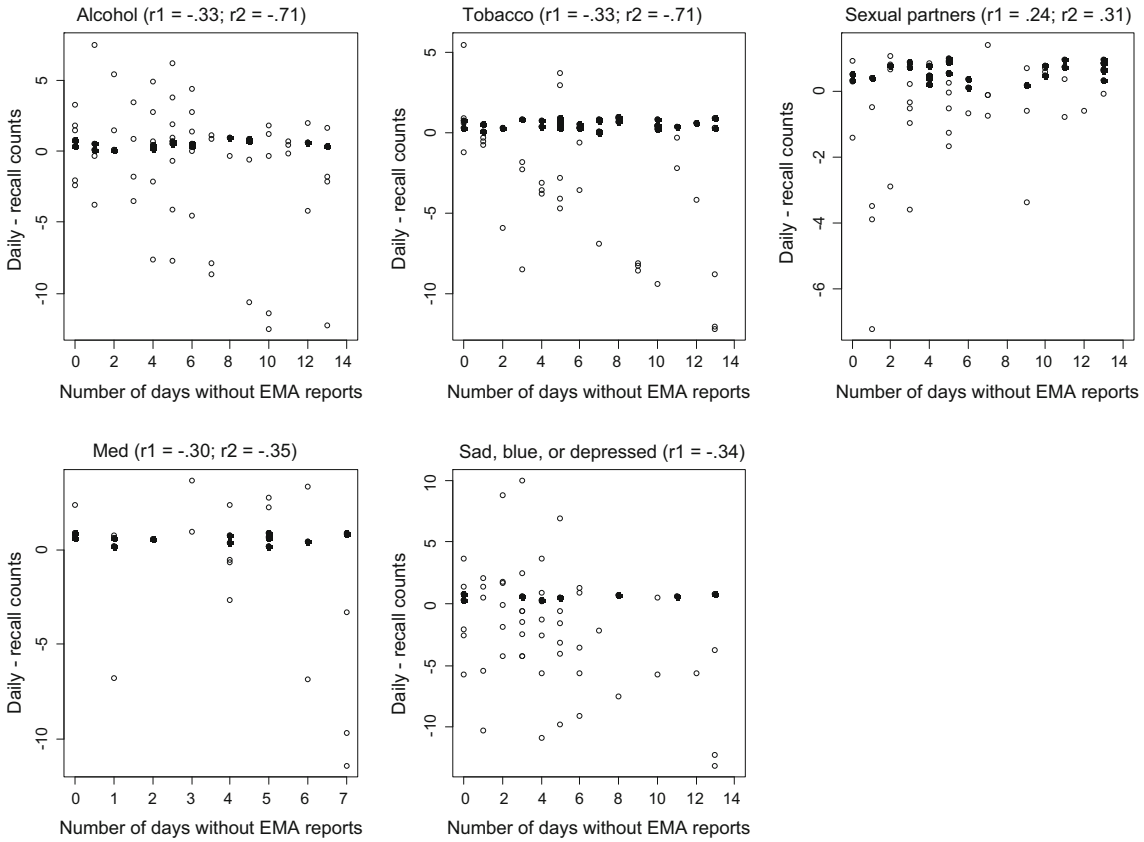

Fig. 2 Plots of exemplar domains showing correlations of the differences between daily and recall reports by the number of days without daily reports. Solid dots represent zero-reports for both daily and recall reports (e.g. abstinent). Pearson correlation coefficients shown for 14-day periods including zero reports (r1) and excluding zero reports (r2)

ATOD Use

Tables 1 and 2 show relatively high correlations when including zero reports (r = 0.65–0.94) and modest correlations for users only (r = 0.14–0.53). Daily-recall comparisons show consistency for almost all domains, indicated by similar mean counts or proportions (differences not statistically significant) and by fairly balanced numbers of discordant pairs with higher and lower daily versus recall reports. Two exceptions are noted, alcohol and tobacco, also shown in Fig. 1. Mean count comparisons for tobacco (shown in Table 1) are higher for recall than daily reports, reflecting previously cited rounding up errors and inconsistent daily reporting. Mean percent comparisons (Table 2) demonstrate consistent reporting of tobacco use, which tends to be either daily or not. Alcohol shows the opposite trend, with significant differences in mean percent of days and discordant pair comparisons both showing higher reports for daily compared to recall responses (Table 2). This trend can be explained by the more episodic nature of alcohol consumption in which event-based reporting (assumed in mean count comparisons) is likely.

Sexual Behaviors

Correlations between daily and recall reports of sexual behaviors show modest correlations when including abstainers (r = 0.54–0.61) for all variables except unprotected sex with HIV-negative or unknown status partners (r = −0.44 for sexually active, but result is driven by low reports and a few outlier points). Discordant pair comparisons show consistent under-reporting in daily reports compared to recall. There were also significantly fewer partners reported daily versus recall (mean of .4 vs. 1.3 partners; $t = -3.76$, $df = 61$, $p < 0.01$) and in contrast to significantly greater number of unprotected sex acts reported daily (mean of 1.8 in daily vs. 1.3 in recall; $t = 2.15$, $df = 70$, $p = 0.03$). Figure 1 illustrates the consistent under reporting for sexual partners as an example.

Reporting Biases and Missing Data

Correlations were also examined for differences between daily reports and recall counts versus the number of missing days of daily reports. Figure 2 shows plots of

1. *Examine multiple indicators of agreement or concordance for reliability and validity.*
2. *Analyze data under complementary reporting patterns (event- and time-based) and missing data assumptions.*
3. *Event-based reporting should be anticipated in future research and application design.*
4. *Short-term recalls at bi-weekly or weekly intervals may be good-enough and more sustainable over time, for easy to recall events and behaviors.*
5. *Predictive validity for change detection may be more important than absolute accuracy and consistency of self-monitoring.*

Fig. 3 Key Recommendations for mHealth Self-monitoring Assessment for PLH

results for the five sample domains in Fig. 1. Negative correlations for alcohol and tobacco and the trend of points below the 0 axis show increasing daily-recall discrepancies as number of missing daily reports increase, which suggests that participants tended to report on use days (i.e., event-based reporting). A similar trend in negative correlations is evident for depressive symptoms and medication adherence. Positive correlations, evident in the sexual partners example in Fig. 2, are more difficult to interpret because there are a number of missing-data scenarios that can explain the trend. In this case, the most likely explanation is non-response in daily survey prompts to repeat the survey for each sexual encounter.

Limitations

There are several limitations in this data that have implications for interpreting results and future work. First, the sample is a convenience sample recruited from an AIDS service organization, and with relatively precise eligibility criteria. Although the sample is diverse in ethnicity, education, socio-economic status, and gender and sexual identity, it may not be representative of all PLH. In addition, much larger samples are needed to untangle multiple potential biases such as recall errors, social desirability, under- or over-reporting, or missing reports. Our analyses, and previous research, use typical statistical comparisons where the null hypothesis is that there is no difference and proof is to show a difference between measurement methods. Therefore, a lack of significance leaves possibility that daily diary and recall are similar or that sample size is too small to make a determination. Future research needs to reverse hypotheses to prove diary and recall techniques are not different, as is done for equivalency trials in pharmaceutical studies of name-brand and generic drugs.

There are also some limitations related to reporting differences between web and smartphone survey questions. Web surveys referred to "Uppers/stimulants like speed, crystal, ice" as a drug category, and mobile surveys referred to "Methamphetamine (Crystal)." Both categories are treated equivalently, which is appropriate in context of Los Angeles with this population. Web surveys also gave separate categories for 'cocaine' and 'crack' use, while mobile surveys had a single category for 'crack-cocaine' use and so we are only able to examine crack use in these analyses. Recall surveys for sexual behaviors were partner-centered (i.e., start with partner, then act, then characteristics of act) while mobile surveys were encounter-centered (i.e., act, then partner, then characteristics of act with partner). The mobile survey required participants to re-open the sexual behavior survey for each partner, which likely compounded the anticipated reticence to report detailed sexual behaviors and participant specific information. In addition, recall reports were based on the five most recent sex partners and so were possibly capped for some participants (5 participants reported more than 5 recent partners at baseline). Finally, the timeframe for phone surveys of substance use and sexual behaviors queried "since last report" to anticipate non-daily reporting, but this may have resulted in some underestimation of days of use.

Discussion

This study demonstrates reliability and validity of daily mobile phone self-monitoring for key domains for PLH under relatively naturalistic participation in a non-treatment setting. Most prior studies of reliability and validity of daily reporting methods examine only one or two measures of inter-method agreement, typically correlations. This study presents a complementary set of methodological tools that begin to address the gap in evidence for reliability and validity methods for mobile health applications [2]. The study also suggests a number of recommendations for future analytic methods, assessment frequency and duration, and survey and application design (see Fig. 3).

Multiple indicators of reliability and validity should inform future development, pilot testing, and validation of mobile health measures. In this study, multiple inter-method agreement statistics, analytic approaches, and visual plots demonstrate complementary inferences on potential biases in daily self-monitoring in comparison to recall. Reliability

correlations were modest to good for all domains and consistent with or higher than results from the limited prior research, except for activity limitations and unprotected sexual intercourse. Assessment of validity by comparisons of means and discordant pairs demonstrate reporting bias trends not evident in reliability correlations, which do not indicate direction of potential reporting biases.

Like prior studies, this study demonstrated some potential for recall to over-report frequent habitual behaviors like tobacco use [20], and under-report episodic and socially undesirable behaviors such as unprotected sex [25] and alcohol use [21] (particularly with ART). The more subjective HRQOL domains reported daily with likert-scale responses had modest to low correlations and high numbers of discordant pairs and means differences, which are likely affected by high variance in reports of states and threshold effects for classifications of days experienced based on daily reports [6, 12, 16].

In this and many prior EMA studies, participants were instructed to complete both time- and event-based EMA. Unlike most prior studies, this analysis explicitly addressed the high potential for some degree of inconsistent daily reporting. Consistent patterns of event-based reporting are indicated when high reliability and concordance rates are exhibited even when daily reporting compliance rates are low (e.g., ATOD use and adherence reported on 50 % of days). Even when only time-prompted instructions are given, participants may be more likely to respond when events occur. Event-based reporting should be anticipated in future application design. The results of this study also suggest that event-based reporting is a reasonable approach for reporting rare and/or highly salient events and can minimize assessment burden.

Alternatively, short-term recalls at bi-weekly or weekly intervals may be good-enough and more sustainable over time for easy to recall events and behaviors. The results of this study demonstrate how participants naturally select reporting frequencies matched to the variability and frequency of self-monitored domains. Future designs should consider these factors in assessment design and not be bound by daily reporting norms for every domain of interest. This is particularly important for applied self-management intervention aims in clinical settings where sustaining self-management activities over several months between clinic visits may be desirable. The authors are currently testing use of ongoing and brief weekly assessments by text-message or interactive voice response in studies with PLH.

The phone surveys for sexual behaviors and ATOD use in this study were also designed with non-daily reporting and short-term recall in mind by querying behaviors since last report. More sophisticated approaches, such as time-stamped backfilling of detailed time-based surveys, can result in more complete data but does not reduce survey burden and could accumulate into a barrier to re-engaging with the self-monitoring activities after a hiatus. Again, daily reporting may not be necessary or feasible for all domains over extended periods.

Finally, predictive validity for change detection may be more important than absolute accuracy and consistency of intensive self-monitoring for self-management applications. While accuracy and consistency (i.e., inter-method reliability and validity) of mHealth assessments are important to assess and understand [2], predictive validity (i.e., ability to detect events or pattern changes) may be paramount for intervention applications.

Conclusion

Mobile phones are offering unprecedented opportunities to massively scale tools based on continuous self-monitoring, which can be applied for research and interventions. The results of this analysis contribute to an emerging evidence-base on reliability, validity, and potential biases of in-the-moment assessments enabled by mobile phones. More work is needed to realize the full potential of such mobile health tools.

Acknowledgments This work was supported by the Center for HIV Identification, Prevention, and Treatment (CHIPTS) NIMH Grant MH58107; and also by the UCLA Center for AIDS Research (CFAR) Grant 5P30AI028697; and the *National Center for Advancing Translational Sciences* through UCLA CSTI Grant UL1TR000124. Comulada's time was also supported by NIMH Grant K01MH089270. Swendeman's time also supported by a career development Grant from the William T. Grant Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Conflict of interest The authors declare that they have no competing interests.

References

1. Swendeman D, Rotheram-Borus MJ. Innovation in sexually transmitted disease and HIV prevention: internet and mobile phone delivery vehicles for global diffusion. Curr Opin Psychiatry. 2010;23:139–44.
2. Kumar S, Nilsen WJ, Abernethy A, Atienza A, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med. 2013;45(2):228–36.
3. Csikszentmihalyi M, Larson R. Validity and reliability of the experience-sampling method. J Nerv Ment Dis. 1987; 175(9):526–36.
4. Shiffman S, Stone AA. Introduction to the special section: ecological momentary assessment in health psychology. Health Psycho. 1998;17(1):3.
5. Fahrenberg J, Myrtek M, Pawlik K, Perrez M. Ambulatory assessment-monitoring behavior in daily life settings. Eur J Psychol Assess. 2007;23(4):206–13.

6. Ebner-Priemer UW, Trull TJ. Ecological momentary assessment of mood disorders and mood dysregulation. Psychol Assess. 2009;21(4):463–75.
7. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol. 2010;15(Pt 1):1–39.
8. Reddy S, Shilton K, Burke J, Estrin D, Hansen M, Srivastava M. Evaluating participation and performance in participatory sensing. In International workshop on urban, community, and social applications of networked sensing systems. Raleigh, North Carolina: 2013.p. 1–5.
9. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92.
10. Centers for Disease Control and Prevention. Health-related quality of life (HRQOL). http://www.cdc.gov/hrqol/. Accessed August 21 2013.
11. Buysse DJ, Ancoli-Israel S, Edinger JD, Lichstein KL, Morin CM. Recommendations for a standard research assessment of insomnia. Sleep: J Sleep and Sleep Disorders Res. 2006;29(9):1155–73.
12. Peters ML, Sorbi MJ, Kruise DA, Kerssens JJ, Verhaak PF, Bensing JM. Electronic diary assessment of pain, disability and psychological adaptation in patients differing in duration of pain. Pain. 2000;84(2):181–92.
13. Mustanski BS. Are sexual partners met online associated with HIV/STI risk behaviours? retrospective and daily diary data in conflict. AIDS Care. 2007;19(6):822–7.
14. Grov C, Golub SA, Mustanski B, Parsons JT. Sexual compulsivity, state affect, and sexual risk behavior in a daily diary study of gay and bisexual men. Psychol Addict Behav. 2010;24(3):487–97.
15. Dockray S, Grant N, Stone AA, Kahneman D, Wardle J, Steptoe A. A comparison of affect ratings obtained with ecological momentary assessment and the day reconstruction method. Soc Indic Res. 2010;99(2):269–83.
16. Solhan MB, Trull TJ, Jahng S, Wood PK. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. Psychol Assess. 2009;21(3):425.
17. Kerr T, Walsh J, Lloyd-Smith E, Wood E. Measuring adherence to highly active antiretroviral therapy: implications for research and practice. Curr HIV/AIDS Rep. 2005;2:200–5.
18. Simoni JM, Kurth AE, Pearson CR, Pantalone DW, Merrill JO, Frick PA. Self-report measures of antiretroviral therapy adherence: a review with recommendations for HIV research and clinical management. AIDS Beh. 2006;10(3):227–45.
19. Haberer JE, Kiwanuka J, Nansera D, et al. Real-time adherence monitoring of antiretroviral therapy among hiv-infected adults and children in rural uganda. AIDS. 2013;27(13):2166–8.
20. Shiffman S. Ecological momentary assessment (EMA) in studies of substance use. Eur J Psychol Asess. 2009;21(4):486–97.
21. Searles JS, Helzer JE, Rose GL, Badger GJ. Concurrent and retrospective reports of alcohol consumption across 30, 90 and 366 days: interactive voice response compared with the timeline follow back. J Stud Alcohol Drugs. 2002;63(3):352.
22. Hopper JW, Su Z, Looby AR, et al. Incidence and patterns of polydrug use and craving for ecstasy in regular ecstasy users: an ecological momentary assessment study. Drug Alcohol Depen. 2006;85(3):221–35.
23. Coxon AP. Parallel accounts? discrepancies between self-report (diary) and recall (questionnaire) measures of the same sexual behaviour. AIDS Care. 1999;11(2):221–34.
24. Garry M, Sharman SJ, Feldman J, Marlatt GA, Loftus EF. Examining memory for heterosexual college students' sexual experiences using an electronic mail diary. Health Psycho. 2002;21(6):629–34.
25. Leigh BC, Gillmore MR, Morrison DM. Comparison of diary and retrospective measures for recording alcohol consumption and sexual activity. J Clin Epidemiol. 1998;51(2):119–27.
26. Shrier LA, Shih MC, Beardslee WR. Affect and sexual behavior in adolescents: a review of the literature and comparison of momentary sampling with diary and retrospective self-report methods of measurement. Pediatrics. 2005;115(5):e573–81.
27. Durant LE, Carey MP. Self-administered questionnaires versus face-to-face interviews in assessing sexual behavior in young women. Arch Sex Behav. 2000;29(4):309–22.
28. Graham CA, Catania JA, Brand R, Duong T, Canchola JA. Recalling sexual behavior: a methodological analysis of memory recall bias via interview using the diary as the gold standard. J Sex Res. 2003;40(4):325–32.
29. Jaccard J, McDonald R, Wan CK, Guilamo-Ramos V, Dittus P, Quinlan S. Recalling sexual partners: the accuracy of self-reports. J Health Psychol. 2004;9(6):699–712.
30. McAuliffe TL, DiFranceisco W, Reed BR. Effects of question format and collection mode on the accuracy of retrospective surveys of health risk behavior: a comparison with daily sexual activity diaries. Health Psycho. 2007;26(1):60.
31. Horvath KJ, Beadnell B, Bowen AM. A daily web diary of the sexual experiences of men who have sex with men: comparisons with a retrospective recall survey. AIDS Beh. 2007;11(4):537–48.
32. Ramanathan N, Swendeman D, Comulada WS, Estrin D, Rotheram-Borus MJ. Identifying preferences for mobile health applications for self-monitoring and self-management: focus group findings from HIV-positive persons and young mothers. Int J Medical Inform. 2013;82(4):e38–46.
33. Swendeman D, Mendenhall B, Lazar M, et al. Daily mobile phone self-monitoring of medication use, mood, stress, sexual behaviours and substance use is feasible and acceptable among people living with HIV (PLHIV) and demonstrates benefits for self-management of health behaviours and quality of life. 19th International AIDS Conference. 2013.
34. Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. Trials. 2009;10:37.
35. Chesney M, Ickovics J. ACTG Adherence baseline questionnaire. San Francisco: AIDS clinical trial group recruitments, adherence and retention sub Committee; 1997.
36. Comulada SW, Weiss RE, Cumberland W, Rotheram-Borus MJ. Reductions in drug use among young people living with HIV. Am J Drug Alcohol Abuse. 2007;33(3):493–501.
37. NIMH Multisite HIV Prevention Trial. Conceptual basis and procedures for the intervention in a multisite HIV prevention trial. AIDS. 1997;11:S29–35.
38. Carney MA, Tennen H, Affleck G, Del Boca FK, Kranzler HR. Levels and patterns of alcohol consumption using timeline follow-back, daily diaries and real-time "electronic interviews". J Stud Alcohol. 1998;59:447–54.
39. SAS Institute Inc, Cary, N.C: SAS Institute Inc.
40. R Development Core Team R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. 2008. http://www.R-project.org. Accessed August 21 2013.