The short-term reliability and long-term stability of salivary immune markers

Grant S. Shields[a],*, George M. Slavich[b], Greg Perlman[c], Daniel N. Klein[d], Roman Kotov[c,d,e],*

[a] Center for Neuroscience and Department of Psychology, University of California, Davis, CA, USA
[b] Cousins Center for Psychoneuroimmunology and Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, CA, USA
[c] Department of Psychiatry, Stony Brook University, Stony Brook, New York, USA
[d] Department of Psychology, Stony Brook University, Stony Brook, New York, USA
[e] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York, USA

ARTICLE INFO

ABSTRACT

Salivary markers of immune function are increasingly commonly used in studies of human health. Yet, few studies have examined the short-term or long-term reliability or stability of these biomarkers, making their measurement properties unclear. We addressed this issue in the present study by collecting two saliva samples, two hours apart, from 426 adolescent girls during a baseline laboratory visit. Then, eighteen months later, we collected the same samples again from a subset of these participants ($n = 113$). The correlations between the two samples collected at each session were generally high (mean $r = 0.67$). In contrast, although single saliva samples were only weakly correlated across 18 month (mean rs = 0.18), averaging the two quantifications within a session considerably improved the reliability (mean $r = 0.27$). In short, salivary immune markers exhibited strong short-term test-retest correlations, and averaging across multiple assessments notably improved long-term test-retest correlations. Additional research is needed to establish the health relevance and mechanisms underlying these potentially useful, non-invasive biomarkers.

1. Introduction

Substantial research has shown that psychosocial factors can influence various components of the immune system (for reviews, see Dantzer and Kelley, 2007; Marsland et al., 2017; Slavich and Irwin, 2014). These findings have in turn prompted numerous biobehavioral researchers to assess immune function, often focusing on inflammatory biomarkers. Inflammatory biomarkers are frequently examined in blood, which is generally regarded as the gold standard method for assessing inflammatory activity. However, there are also drawbacks to this approach. For example, some individuals refuse to provide blood samples. Additionally, it can be difficult or impossible to collect blood samples in some study designs (e.g., at-home studies).

An increasingly common alternative to collecting blood involves measuring inflammatory biomarkers in saliva. Salivary inflammatory biomarkers are believed to reflect an interaction between systemic and local immune activity as well as oral hygiene (Slavish et al., 2015). Numerous studies have assessed salivary immune markers in recent years, and this work has suggested that these biomarkers may index important aspects of immune function, such as stress responsivity and biological resource redistribution (e.g., Shields et al., 2016; Slavish et al., 2015). However, few studies have examined the measurement characteristics of salivary inflammatory biomarkers. Therefore, their basic methodological characteristics and measurement properties remain relatively unknown. This is problematic because in order to have utility, a biomarker must be able to be assessed reliably.

Riis and colleagues (2014) examined the reliability of inflammatory biomarkers at three assessment points separated by one year each. They found that intercorrelations among salivary inflammatory biomarkers were high in a baseline sample, but correlations within each inflammatory biomarker from one timepoint to the next (i.e., over a one-year period) were often nonsignificant. This analysis represented a critical first-step in documenting the long-term (un)reliability of salivary inflammatory biomarkers, but because all samples were separated by one year, it remains unclear to what extent these results are due to measurement error versus changes in inflammation over time. Moreover, there are presently no guidelines for designing studies to optimize the reliability of salivary immune markers in psychoneuroimmunology

research.

One strategy that could improve the long-term reliability of salivary inflammatory biomarker assays is to average two or more samples (i.e., versus a single sample, per usual in studies of inflammatory biomarkers). Although most studies assay samples in duplicate, this is only a partial solution. Assaying in duplicate helps address measurement error introduced by the assay and technician; however, unlike taking an average of two independent samples, assaying in duplicate does not help address measurement error introduced by the person who takes the sample, variability in storage procedures between participants (e.g., collection-to-freezing time), or rapid changes in the environment of the mouth.

Psychometric research has long established that single-item measures show poor reliability (Gliem and Gliem, 2003), and averaging two or more samples has been used to improve reliability in prior medical research (Jensen and McFarland, 1993). Moreover, with two samples, it is possible to utilize the Spearman-Brown prophecy formula—a statistical method that projects the reliability of a test if the number of items change—to determine the number of samples required to achieve a desired reliability for each analyte and calculate disattenuated correlations (i.e., correlations correcting for measurement error). Although averaging two samples collapses within-person variance, enhancing the temporal stability of these biomarkers may offer important advantages, such as providing a better ability to predict the onset of depression or distinguish subtypes within a heterogeneous disorder, such as schizophrenia. To date, however, no study has examined whether creating a composite from multiple same-day samples can improve the long-term stability of salivary immune markers.

To address these issues, we recruited a large sample of adolescents and quantified their salivary levels of nine commonly investigated immune markers in two samples (separated by two hours) at both a baseline and follow-up assessment (18 months later). The biomarkers were chosen based on a comprehensive literature review of biobehavioral research and included tumor necrosis factor-α (TNF-α), interleukin (IL)-1β, IL-6, IL-8, IL-10, IL-18, IL-33, monocyte chemoattractant protein 1 (MCP-1), and C-reactive protein (CRP). First, we characterized both the short-term (same-session) reliability and long-term (longitudinal) test-retest correlations of the analytes. Second, we examined whether creating composite values from the two samples within each session improved the long-term test-retest correlations of the analytes. Finally, we calculated disattenuated correlations to determine the stability of these salivary immune markers over 18 months, correcting for measurement error.

2. Method

2.1. Participants

Participants were 426 adolescent girls who completed Wave 3 of the Adolescent Development of Emotions and Personality Traits (ADEPT) project (M_{age} = 15.84 years old; SD = 0.63). ADEPT is a longitudinal study examining factors affecting female adolescent wellbeing and depression risk. Inclusion criteria for enrollment into ADEPT were English fluency, ability to read and comprehend questionnaires, age between 13.5 and 15.5 years old, and a biological parent consenting to participate in the study. Exclusion criteria were a lifetime history of a major depressive episode (MDE) or dysthymia, or intellectual disability. A diagnosis of an autoimmune disorder was not part of the exclusion criteria for this study, and approximately 2% (eight participants) reported being diagnosed with an autoimmune disorder by a doctor. Excluding these participants did not alter any reliability or stability outcome.

Salivary immune biomarker data were collected at ADEPT assessments Wave 3 and Wave 5; as such, Wave 3 will hereafter be referred to as "Baseline" and Wave 5 as "Follow-up". Participants were predominately White (81.2%), followed by Hispanic (10.1%), Black

(5.2%), Asian (2.6%), American Indian (0.2%), and other (0.7%). Of this cohort of 426 girls, 113 were also assayed at the follow-up assessment (i.e., Wave 5) 18 months later. The retention rate for the larger cohort was over 90%, but we assayed only a randomly sampled subset of these participants due to limited funding. Participants whose samples were randomly selected to be assayed did not differ from unselected participants with respect to age, body mass index (BMI), parent years of education, race, or oral health at Follow-up, ps > 0.268.

2.2. Materials and procedure

Participants provided two saliva samples—120 min apart—at both the baseline assessment (i.e., Baseline Sample 1 and Baseline Sample 2) and follow-up assessment that occurred 18 months later (i.e., Follow-up Sample 1 and Follow-up Sample 2). Nearly all saliva samples were provided between 3 pm and 8 pm; participants who were unable to attend the study during that time at Baseline were assessed at approximately the same time during the Follow-up assessment when possible. After providing the first saliva sample (i.e., Sample 1), participants remained in the lab for 120 min, during which time they completed unrelated measures—none of which were inherently stressful. Participants were not allowed to eat anything during this time. After 120 min had elapsed, participants provided the second saliva sample (i.e., Sample 2). Saliva was collected via passive drool and immediately stored in a −80 °C freezer until batch assayed at the UNC Cytokine and Biomarker Analysis Facility.

2.2.1. Assays

Salivary levels of inflammatory biomarkers were determined using multiplex immunoassay kits purchased from R&D Systems (Minneapolis, MN) with a Bio-Plex 200 (Luminex) instrument. Assays were conducted following manufacturer instructions. The mean fluorescence intra-assay coefficient of variation (CV) was 2.99%, inter-assay CV was 10.27%, and the average percent of observed to expected values of known concentration was 99.7%. All values are given in pg/mL.

2.2.2. Oral health

Participants completed an interview examining oral health/hygiene at each assessment (see Supplemental Material).

2.3. Data analysis

Pearson correlations and Spearman-Browne reliabilities were used in analyses. Additional information on the analytic strategy is available in the Supplemental Material.

3. Results

3.1. Detection rates

Detection rates were very good for nearly all analytes. Analytes with a detection rate below 80% in either baseline sample were IL-10 and IL-33. These detection rates were essentially equivalent or better at the follow-up assessment (see Table 1). Due to poor detection rates, we do not consider IL-10 and IL-33 further.

3.2. Short-term reliability

Descriptive statistics, differences between means, and correlations between samples for each analyte (derived from the two samples taken two hours apart at each of the assessments) are presented in Table 1. Correlations between Baseline Sample 1 and Baseline Sample 2 were strong, rs > 0.50, ps < 0.001. Correlations between Follow-up Sample 1 and Follow-up Sample 2 were essentially equivalent, rs > 0.44, ps < 0.001. At Baseline, the mean short-term reliability (i.e., the average test-retest correlation between each analyte's Sample 1 and Sample 2) was r = 0.67; at

Table 1
Descriptive Statistics, Mean Differences, and Correlations of Each Sample for Each Analyte at Each Assessment.

	Sample 1				Sample 2				Analyses			
	N	Detection Rate	*Mean*	*SD*	N	Detection Rate	*Mean*	*SD*	Cohen's *d*	Paired *N*	*r*	Reliability of Sum
Baseline												
TNF-α	417	97.9%	3.11	4.42	414	97.2%	3.37	3.37	**0.13**	391	**0.65**	0.79
IL-1β	426	100.0%	137.26	174.72	416	97.7%	191.14	233.36	**0.36**	403	**0.70**	0.82
IL-6	422	99.1%	4.18	8.76	413	96.9%	4.33	8.76	0.01	401	**0.72**	0.84
IL-8	413	96.9%	528.55	401.61	389	91.3%	711.27	437.34	**0.51**	379	**0.51**	0.67
IL-18	378	88.7%	7.74	12.95	385	90.4%	8.65	14.26	**0.08**	344	**0.55**	0.71
CRP	405	95.1%	112.98	269.43	393	92.3%	128.58	308.95	**0.02**	381	**0.81**	0.90
MCP-1	422	99.1%	374.17	724.25	415	97.4%	389.62	651.22	0.07	403	**0.64**	0.78
Poor Detection												
IL-10	159	37.3%	1.08	0.98	176	41.3%	1.13	1.07	0.06	97	**0.25**	0.41
IL-33	247	58.0%	1.36	1.45	270	63.4%	1.30	1.41	−0.01	177	**0.45**	0.62
Follow-up												
TNF-α	113	100.0%	8.17	3.90	113	100.0%	8.42	4.39	−0.04	110	**0.54**	0.70
IL-1β	113	100.0%	235.76	264.46	113	100.0%	280.48	297.45	**0.17**	108	**0.61**	0.76
IL-6	108	95.6%	8.97	14.75	111	98.2%	9.00	18.69	−0.08	104	**0.66**	0.79
IL-8	112	99.1%	792.49	601.28	109	96.5%	947.75	642.20	**0.26**	106	**0.39**	0.56
IL-18	102	90.3%	14.13	17.89	105	92.9%	15.15	15.80	0.16	91	**0.59**	0.74
CRP	113	100.0%	153.76	419.41	113	100.0%	230.37	842.64	0.03	111	**0.72**	0.84
MCP-1	113	100.0%	288.29	270.43	113	100.0%	359.31	386.78	**0.18**	108	**0.49**	0.66

Note: For analyses, values greater than three SDs ± the mean were removed. A positive Cohen's *d* indicates greater values at Sample 2. Significant (*p* < .05) differences and/or correlations are indicated by boldface font.

Table 2
Short-Term (Same Session) Estimated Reliability of Analyte by Number of Samples.

Number of Samples	1	2	3	4
Analyte			Reliability	
TNF-α	0.65	0.79	**0.85**	**0.88**
IL-1β	0.70	**0.82**	**0.87**	**0.90**
IL-6	0.72	**0.84**	**0.88**	**0.91**
IL-8	0.51	0.67	0.76	**0.81**
IL-18	0.55	0.71	0.79	**0.83**
CRP	**0.81**	**0.90**	**0.93**	**0.95**
MCP	0.64	0.78	**0.84**	**0.88**

Note: Estimated short-term reliability calculated by the Spearman-Brown formula from same-session samples (i.e., separated by two hours). Reliabilities > 0.80 are bolded.

Follow-up, the mean short-term reliability was *r* = 0.58. We used Spearman-Brown prophecy formula to estimate the number of samples per person needed to achieve reliability of 0.80 for each analyte, and it ranged from two to four (see Table 2).

As for within-person changes, nearly all of the immune markers exhibited negligible within-person changes between Sample 1 and Sample 2 at both assessments (i.e., |*d*|s < 0.14). The primary exceptions were IL-1β and IL-8, which increased from Baseline Sample 1 to Baseline Sample 2 (*d* = 0.36 and *d* = 0.51, respectively) and to a lesser extent from Follow-up Sample 1 to Follow-up Sample 2 (*d* = 0.17 and *d* = 0.26, respectively). In sum, most of the salivary inflammatory biomarkers assessed showed significant within- and between-person reliability and negligible changes over a two-hour period.

3.3. Long-term test-retest correlations

Next, we examined the test-retest correlations of these salivary immune markers over 18 months. We first correlated individual samples—namely, Baseline Sample 1 with Follow-up Sample 1, and Baseline Sample 2 with Follow-up Sample 2. All of the correlations were positive, ranging from 0.04 to 0.32 (mean *r*s = 0.18 for both Baseline Sample 1 with Follow-up Sample 1 and Baseline Sample 2 with Follow-up Sample 2, *p*s > 0.05), and five of the fourteen correlations were significant (see Table 3).

Table 3
Correlations Within Analytes Between Each Assessment (i.e., Baseline and Follow-up).

	Correlation Between Baseline and Follow-up				
Analyte	Sample 1	Sample 2	Composite	Partial	Disattenuated
TNF-α	0.07	0.17[†]	0.22*	0.22*	0.30**
IL-1β	0.18[†]	0.19*	0.30**	0.26**	0.38***
IL-6	0.04	0.19[†]	0.10	0.12	0.13
IL-8	0.17[†]	0.23*	0.27**	0.27**	0.44***
IL-18	0.18[†]	0.11	0.37***	0.30***	0.51***
CRP	0.31***	0.20*	0.31***	0.30***	0.36***
MCP	0.32***	0.18[†]	0.31**	0.34***	0.43***

Note: [†]*p* < .10, *p < .05, **p < .01, ***p < .001. Columns provide the correlation between Baseline and Follow-up for the given correlation (e.g., Baseline Sample 1 with Follow-up Sample 1). Composite represents the correlation between variables after averaging both samples at each assessment. Partial represents the correlation between the composite variables after covarying baseline and follow-up oral health as well as baseline and follow-up sample collection time. Disattenuated represents the correlation between the composite variables after correcting for measurement error.

3.4. Enhancing long-term test-retest correlations and stability

To attempt to improve the long-term test-retest correlations of these analytes, we created a composite score that averaged the values of Sample 1 and Sample 2 for each analyte at both Baseline and Follow-up. Doing so greatly improved the test-retest correlations of these biomarkers (see Table 3), with correlations now ranging from 0.10 to 0.37 (mean *r* = 0.27, *p* = .004). Moreover, the long-term test-retest correlations of all analytes were significant except for IL-6. Therefore, using a composite score from two samples substantially increases the long-term test-retest correlations of salivary immune markers.

Because oral health/hygiene and sample collection time can both strongly influence salivary inflammatory biomarkers, we conducted analyses of the composite correlation (i.e., averaging Sample 1 and Sample 2 for use at both Baseline and Follow-up) controlling for oral health/hygiene at baseline and follow-up as well as sample collection time at both baseline and follow-up. As shown in Table 3, controlling for these covariates did little to influence this composite: the mean

difference in magnitudes between these correlations was $r = 0.01$, and the largest difference was $r = 0.07$. Therefore, oral health/hygiene and time of sample did not exert strong influences on relative changes in these inflammatory biomarkers in our sample.

Finally, to estimate the 18-month stability of these markers, we corrected the long-term correlations for attenuation (i.e., measurement error). The disattenuated correlations ranged 0.13–0.51 (mean $r = 0.37$), indicating that the long-term stability of these salivary immune markers is moderate, on average, when correcting for measurement error (see Table 3).

Intraclass correlation coefficients (ICCs) are presented within the Supplemental Material.

4. Discussion

Despite growing interest in salivary immune markers, very little is known about their measurement properties. To address this issue, we assessed the short-term reliability of nine salivary markers of inflammation and the long-term stability of seven. Most of these salivary biomarkers (7/9) were highly detectable, with nearly all showing detectability rates of greater than 80%. On average, the short-term reliability of these markers was strong (i.e., mean $r = 0.67$). The long-term test-retest correlations were weaker, but here we show that taking the average of two samples substantially improves the long-term test-retest correlations of salivary immune markers. Inflammatory biomarkers showing the greatest long-term stability were IL-18, IL-8, and MCP; in contrast, IL-6 was very unstable.

One of the most interesting results obtained was that, despite similar short-term reliability, CRP showed relatively higher long-term stability than most salivary immune markers, whereas IL-6 showed relatively lower long-term stability. This difference in relative stability may thus suggest that salivary IL-6 is more sensitive than CRP to state-related factors, such as acute stress. Some research supports this idea, with acute stress effects on CRP being smaller than IL-6 (e.g., Marsland et al., 2017). Future research should examine this potential explanation and other potential reasons to better understand the relative instability of IL-6 relative to CRP.

Although we did not measure serum immune markers, concentrations of the measured salivary immune markers were generally in agreement with prior literature. This research has found that salivary concentrations of these markers are roughly equivalent with serum, with the exceptions of IL-1β and IL-8, which are higher in saliva than serum, and CRP, which is lower in saliva (Byrne et al., 2013; Riis et al., 2014). High salivary concentrations of IL-1β and IL-8 are thought to reflect the importance of neutrophils in oral health, as these cytokines attract and activate neutrophils (Riis et al., 2014); low CRP concentrations are due to the fact that this protein is primarily made in the liver and cannot easily pass into saliva (Byrne et al., 2013). It is also worth speculating that non-local production of CRP may be responsible for its relatively greater stability in this study.

Because we did not record the time each participant took to fill their vials, adjusting for flow rate was not possible. Crucially, however, salivary IL-1β (Salimetrics, 2017a), IL-6 (Izawa et al., 2013), and CRP (Salimetrics, 2017b) are known to be independent of flow rate. To our knowledge, no study has examined whether salivary TNF-α, IL-8, IL-18, or MCP-1 are flow-rate dependent. Importantly, though, flow-rate dependency of these salivary inflammatory biomarkers would lower their reliabilities; because the reliabilities of these analytes was similar to the flow-rate independent analytes, we do not believe that flow-rate adjustment would have substantially altered the results.

More broadly, these results replicate the findings of Riis et al. (2014), who found that salivary inflammatory biomarkers exhibit small-to-moderate test-retest correlations over an 18-month period. They also extend these findings, though, by showing that using a composite of samples obtained at different times within an assessment produces much better long-term test-retest correlations than obtaining only one sample.

These results have several implications for research employing salivary immune markers. For example, they provide evidence that most of the salivary biomarkers being measured are highly detectable and reliably index immunological function. Most importantly, they also suggest that longitudinal studies utilizing salivary measures that are interested in examining changes or differences in, or the stability of, immune function over time would benefit from collecting multiple samples per session.

Several study limitations should be noted. First, the sample was young and female, and additional research is needed to examine the generalizability of these findings to other populations. Second, we examined the results of only one type of assay kit, and it is possible that different kits would yield different results. Third, we did not test the potential advantages of obtaining more than two samples per session or of shortening the interval between samples, both of which could further enhance the reliability estimates. Fourth, we did not assess these biomarkers in blood, so we could not compare the stability of these markers in saliva to blood. Finally, this study was not designed to identify factors or processes that could have influenced changes in immune function between the baseline and follow-up assessments, such as diet, sleep, stress, and health behaviors.

In conclusion, although serum-based immune markers have played a major role in psychoneuroimmunology and health research to date, salivary inflammatory biomarkers are becoming increasingly used due to their many advantages (e.g., less expensive and invasive, easier to obtain, etc.). Here, we show that despite existing criticism of this sampling technique, salivary immune markers are highly detectable and exhibit an average long-term stability of $r = 0.37$. Moreover, by using a composite of two samples, multiple salivary immune markers—namely, TNF-α, IL-1β, IL-8, IL-18, CRP, and MCP—demonstrate significant test-retest correlations over 18 months, therefore providing evidence of their suitability for use in studies assessing immune function.

Declaration of Competing of Interest

There are no conflicts of interest with respect to the authorship or publication of this article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbi.2019.06.007.

References

Byrne, M.L., O'Brien-Simpson, N.M., Reynolds, E.C., Walsh, K.A., Laughton, K., Waloszek, J.M., Woods, M.J., Trinder, J., Allen, N.B., 2013. Acute phase protein and cytokine levels in serum and saliva: a comparison of detectable levels and correlations in a depressed and healthy adolescent sample. Brain Behav. Immun. 34, 164–175. https://doi.org/10.1016/j.bbi.2013.08.010.

Dantzer, R., Kelley, K.W., 2007. Twenty years of research on cytokine-induced sickness behavior. Brain Behav. Immun. 21, 153–160. https://doi.org/10.1016/j.bbi.2006.09.006.

Gliem, J., Gliem, R.R., 2003. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. In: Midwest Res. to Pract Conf. Adult, Contin. Community Educ, pp. 82–88. https://doi.org/10.1109/PROC.1975.9792.

Izawa, S., Sugaya, N., Kimura, K., Ogawa, N., Yamada, K.C., Shirotsuki, K., Mikami, I., Hirata, K., Nagano, Y., Nomura, S., 2013. An increase in salivary interleukin-6 level following acute psychosocial stress and its biological correlates in healthy young adults. Biol. Psychol. 94, 249–254. https://doi.org/10.1016/j.biopsycho.2013.06.006.

Jensen, M.P., McFarland, C.A., 1993. Increasing the reliability and validity of pain intensity measurement in chronic pain patients. Pain 55, 195–203. https://doi.org/10.1016/0304-3959(93)90148-I.

Marsland, A.L., Walsh, C., Lockwood, K., John-Henderson, N.A., 2017. The effects of acute psychological stress on circulating and stimulated inflammatory markers: a systematic review and meta-analysis. Brain Behav. Immun. 64, 208–219. https://doi.org/10.1016/j.bbi.2017.01.011.

Riis, J.L., Out, D., Dorn, L.D., Beal, S.J., Denson, L.A., Pabst, S., Jaedicke, K., Granger, D.A., 2014. Salivary cytokines in healthy adolescent girls: intercorrelations, stability, and associations with serum cytokines, age, and pubertal stage. Dev. Psychobiol. 56, 797–811. https://doi.org/10.1002/dev.21149.

Salimetrics, 2017a. Salivary IL-1β ELISA kit.

Salimetrics, 2017b. Salivary C-reactive protein ELISA kit. https://doi.org/10.1074/jbc.R400025200.

Shields, G.S., Kuchenbecker, S.Y., Pressman, S.D., Sumida, K.D., Slavich, G.M., 2016. Better cognitive control of emotional information is associated with reduced pro-inflammatory cytokine reactivity to emotional stress. Stress 19, 63–68. https://doi.org/10.3109/10253890.2015.1121983.

Slavich, G.M., Irwin, M.R., 2014. From stress to inflammation and major depressive disorder: a social signal transduction theory of depression. Psychol. Bull. 140, 774–815. https://doi.org/10.1037/a0035302.

Slavish, D.C., Graham-Engeland, J.E., Smyth, J.M., Engeland, C.G., 2015. Salivary markers of inflammation in response to acute stress. Brain Behav. Immun. 44, 253–269. https://doi.org/10.1016/j.bbi.2014.08.008.