ISCTM Comment: IFCN GREENBEAN Checklist and Explanation & Elaboration for Reporting Studies Evaluating the Effectiveness of EEG-Based Biomarkers

5 November 2024

To: Joshua Ewen and Sándor Beniczky

The International Society for CNS Clinical Trials and Methodology (ISCTM) welcomes this opportunity to respond to the IFCN request for comment regarding the reporting standards: *The GREENBEAN Checklist for Reporting Studies Evaluating the Effectiveness of EEG-Based Biomarkers* and *The GREENBEAN Checklist for Reporting Studies Evaluating the Effectiveness of EEG-Based Biomarkers: Explanation & Elaboration.*

The ISCTM offers these comments for consideration based on our experience and expertise in human CNS research. The ISCTM is an independent organization focused on advancing the development of improved treatments for CNS disorders.

The ISCTM formed a group, led by Kemi Olugemo and Atul Mahableshwarkar, to review and provide comments on behalf of the Society. Working Group members consist of scientists, clinicians, trialists, statisticians and drug developers from both industry and academia. No member of this Working Group received compensation for comments provided. Importantly, comments represent individual opinions and not that of the institution, agency, or company affiliation of group members.

The authors (in alphabetical order) of the comments provided below are:

Michael Avissar, MD, PhD, *Alto Neuroscience*
Kjartan Frisch Herrik, PhD, *H. Lundbeck A/S*
Seth C. Hopkins, PhD, *Otsuka Pharmaceutical Development & Commercialization*
Bruce J. Kinon, MD, *Biopharmaceutical Consultant*
Kristiina Kompus, PhD, *H. Lundbeck A/S*
Atul R. Mahableshwarkar, MD, DLFAPA, *Cybin* (co-chair)
David Matthews, PhD, *Beacon Biosignals*
Dona Murphey, MD, PhD, *Synthesys Brain Health (Lifelines Neuro)*
Kemi Olugemo, MD, FAAN, *Korro Bio* (Chair)
Aneeta Saxena, MD, *Biogen*
Leif Simmatis, PhD, *University of Toronto; Cove Neurosciences, Inc.*

COMMENTS ON THE GREENBEAN CHECKLIST FOR REPORTING STUDIES EVALUATING THE EFFECTIVENESS OF EEG-BASED BIOMARKERS:

General Comments

We applaud the IFCN for their efforts in generating the GREENBEAN Checklist and accompanying Explanation & Elaboration document for Reporting Studies Evaluating the Effectiveness of EEG-Based Biomarkers. ISCTM recommends replacing the term "effectiveness" with "utility," as this is felt to more accurately capture the essence of the checklist & explanation/elaboration document. We would recommend these documents are published as a single document.

An important criterion for a usable EEG biomarker is the test-retest reliability, which can set a ceiling on its usefulness. We therefore propose adding a paragraph to the introduction about reporting known test-retest reliability of proposed EEG marker(s).

For trials involving auditory evoked potentials, subjects should be tested for hearing deficits at Screening, including ability to detect appropriate tone and volume relevant to concept of use. For sites with limited access to audiometry, one could also screen with medical history. The same comment applies to visual evoked potentials depending on COU and staging of trial.

Time-frequency analysis are becoming increasingly recognized as potentially more informative and reliable than time domain ERP measures. Setting standards for reporting on it will ensure better comparability across studies.

Specific Comments

GREENBEAN Checklist

	Line number(s) of the relevant text *(e.g. 20-23)*	Comment and rationale	Proposed text
1	23	Clarify that the implementation of EEG has been successful/unsuccessful, not the technology itself	“While particular approaches to EEG may be new, EEG has been implemented (both successfully and unsuccessfully) for nearly a…”
2	30	It is unclear what "little cross-talk" is alluding to and potentially implies that EEG developed in its own silo.	Recommend rephrasing or deleting "but with little cross talk. “Or state "Independently rather than "in parallel."
3	43-44	Clarification	Potential to remove parenthesis
4	64-78	For the COU section, we recommend linking it to the COU discussions unfolding in the bioanalytical community, to help integrate the validation approaches to EEG biomarkers with fluid biomarkers. See e.g., Cowan, K. J., Kunz, U., Blattmann, P., Gulati, P., Hughes, R., Andersen, L., ... & Timmerman, P. (2024). A European Bioanalysis Forum recommendation for requiring a context-of-use statement for successful development and validation of biomarker assays. Bioanalysis, 1-8.	

5	67	Clarify monitoring for characterizing disease progression vs response to interventions	
6	68-72	Validation seems to be divorced from other biomarker frameworks and processes.	To encompass not only the specific technological characteristics of EEG but also the range of COUs for which EEG-based biomarker are used, we developed a new guideline rather than an extension to STARD or other biomarker development and evaluation frameworks, under the auspices of the International Federation of Clinical Neurophysiology (IFCN) and in collaboration with the EQUATOR (Enhancing the QUAlity Of health Research) Network. Titled "Guidelines for Reporting EEg/Neurophysiological Biomarker Evaluation for Application to Neurology and neuropsychiatry" (GREENBEAN), these reporting guidelines are intended to fill a scope beyond the diagnosis-only focus of STARD (Bossuyt et al., 2015), the specific-condition-focus of other guidelines (Webb et al., 2015), ~~and the~~ specific-EEG-analysis-type focus of ~~yet~~ others (Picton et al., 2000), and the digital health focus of yet others (Goldsack et al., 2020). As a concrete example, classical biomarker concepts like analytical validation are uniquely challenging when a reference standard cannot be easily ascertained for a novel method, as in the

			case of signals derived from brain electrical activity. Thus, new steps for biomarker development and evaluation should be considered in the context of EEG to reflect these challenges.
7	71	As this is the first use, please spell out acronym STARD	
8	83	Most epilepsy centers don't have access to Wada. Not typical test to predict changes to epilepsy surgery. The Wada test remains one of the only tests that can predict the impact of surgery on memory, however fMRI, and Neuropsychological testing are used more routinely to predict memory outcomes in surgical epilepsy.	For example, the Wada test can be considered a standard of care for prediction of memory outcomes in surgical epilepsy patients.
9	114-157	It is not clear where the concept of “phases” originates from, and it may create confusion in clinical trial context, where Phase 1, Phase 2 etc. usually are interpreted as certain types of interventional trials	Consider if “Stage” could better convey the concept.
10	115-118	Other existing frameworks should be aligned with this one to avoid causing confusion. E.g., for V3/+, they have multiple steps that correspond to "clinical validation" in some capacity	Historically biomarker validation seems to evolve in steps, and different subsets of reporting-guideline items would be relevant for studies on different steps of this process and may also correspond to different stages of frameworks developed in other domains/applications such as the V3/V3+ frameworks for general digital health technologies.

11	119-125	There needs to be consideration of the possible effects that analytical pipelines may have on outputs. E.g., different analyses of a given datapoint may lead to vastly different outputs, which may influence the validity of the putative biomarker	These are exploratory studies that report a statistical relationship between the EEG metric and some aspect of the clinical state or diagnosis. They may include case-control comparisons, two-group comparisons, correlation between a physiological (EEG) measure and a clinical variable, data-driven cluster identification, or identification of EEG measure in which clinical group is in tail of normative distribution. Particularly for prognostic, predictive, monitoring and risk COUs, these studies may also include longitudinal data collection in which many EEG variables are explored. The precise steps required to compute the biomarker may also be provided given the potential impact of EEG analytical processes on outputs.
12	120	Clarification to be more patient-centric	relationship between the EEG metric and some aspect of the clinical state or diagnosis that is meaningful to the patient and/or caregivers
13	122	Either remove "individual-patient level" or remove "cross-validation…test-retest reliability" etc. These are more or less mutually exclusive concepts	
14	122	Clarification re: contrasting internal "test sets" from e.g., a single site (P2) vs external site data (P3+)	They often use independent “test” datasets (in contrast to test datasets derived from the same set as training/validation data as in Phase II) to estimate out-of-sample accuracy measures (including

			sensitivity/specificity, positive and negative predictive values for binary scale-of-measurement; confidence intervals or *R*2 for continuous scale-of-measurement).
15	122	More concise language to standardize and align with above steps	Phase 4C: External clinical and demographic generalization
16	122	Expand beyond just age	also performs to a clinically useful degree across demographic groups
17	122		Consider adding the following: • Timing of neurophysiological data in relation to disease state (before, at, or after symptom onset and severity of disease when data was collected). • State-based vs state independence (does the biomarker require collection in a specific wake-sleep state or is it independent of state). • Data collection environment (research data collected with cap vs clinical study applied by technologist). • Demographics (including sex, race/ethnicity, and age) of population studied (ensuring appropriate understanding of the COU).
18	173	The context and methods for collecting quantitative EEG data are as important as how the data are analyzed.	For purposes of comparison across studies, we recommend characterization of the data collection setup (i.e. dry electrode cap versus traditional 10:20),

			the subjects (i.e. demographics including at least sex, race/ethnicity, and age), sleep state, and clinical stage (including preclinical). Whether and how data analyses were fully automated or expert-driven/expert-dependent (for set-up, analytical pipeline, or interpretation), should be reported.

Explanation and Elaboration

	Line number(s) of the relevant text *(e.g. 20-23)*	Comment and rationale	Proposed guidance text
1	224	It is important to maintain blinding in clinical trials if using markers as predictive biomarkers. Readers should be informed of any blinding or lack of it.	Consider adding language about blinding of marker status in clinical trials (blinding from investigators/sites and participants)
2	289	Cannabis and other mind-altering substances are increasing in popularity and may impact EEG and behavior. Reporting on any drug testing would be useful.	Consider adding language about recreational substances and reporting of any urine toxicology during EEG assessment days
3	358	For scalability, EEG techniques that allow self-administration or home use may be advantageous. It might be worth adding a few sentences about such approaches and how they should be reported.	Consider adding language around self-administered or mobile EEG technology
4	373	As a fully spontaneous resting state does not exist, we recommend a statement on what the participant is instructed to do during the resting state.	Consider adding language that encourages discussion of behavioral instruction for spontaneous (resting) EEG

			as it may be relevant (fixation cross, distraction task like counting? Etc.).
5	388-412	A common challenge in academic EEG biomarker discovery papers is that the analytical definition is obscure, or multiple different parameters with different sensitivity and specificity are discussed as if they would be interchangeable.	We recommend this section is edited for even more clarity to help authors to structure their report. The data analysis pipeline reporting needs to cover the following: - software used for data handling after analysis (version, manufacturer, operation system) - all data preprocessing steps in the order they were undertaken - for any steps of automated data cleaning, either refer to module/setting in the used software; if custom code then provide code as supplement to report - for any steps of manual data cleaning, provide training of analysts, which guidelines or principles they were following, any bias mitigation (e.g., double-reading, adjudication) - selection or clustering of electrodes (which electrodes chosen, rationale: prespecified electrodes or based on pre-specified principles of data review, e.g. location of certain peak?) - all signal processing steps need to be described in sufficient detail to allow replication (either specify settings used in the software; for custom code add code to the report; include the mathematical formulae in the report, or refer to sources which contain the formulae)

			- the actual analytical definition of the biomarker which is being validated needs to be unambiguously stated in the report (e.g., "peak amplitude of N1 in muV; peak defined as maximum value in measured in electrode Oz at 200-800 ms post-stimulus, peak calculated as average amplitude value across -20 and +20 ms around peak"; or "phase locking value in the gamma band (38-42 Hz) between the electrodes Fz and Oz").
6	567	Multiple comparison correction may not be as crucial once candidate markers are identified in a discovery set and selected for testing in a replication/test set.	Consider adding language about whether correcting for multiple comparisons is suggested when using a discovery set and replicating in a test set (i. e., statement about how to report statistics if you have an independent replication set)
7	705	Time-frequency measures may be more informative and reliable than time domain ERPs. It would be best to set reporting standards now rather than later to make sure data is reported in a consistent manner on ESRPs.	Consider adding another section or paragraph for reporting on ESRPs (time frequency measures) which are increasingly being used and showing promise. We suggest requiring time-frequency heatmaps with boxes around time frequency regions of interest that are being analyzed.