2014 IEEE International Conference on Systems, Man, and Cybernetics
October 5-8, 2014, San Diego, CA, USA

Identifying Engineering, Clinical and Patient's Metrics for Evaluating and Quantifying Performance of Brain-Machine Interface (BMI) Systems

Jose L. Contreras-Vidal, *Senior Member, IEEE*

***Abstract*— Brain-machine interface (BMI) devices have unparalleled potential to restore functional movement capabilities to stroke, paralyzed and amputee patients. Although BMI systems have achieved success in a handful of investigative studies, translation of closed-loop neuroprosthetic devices from the laboratory to the market is challenged by gaps in the scientific data regarding long-term device reliability and safety, uncertainty in the regulatory, market and reimbursement pathways, lack of metrics for evaluating and quantifying performance in BMI systems, as well as patient-acceptance challenges that impede their fast and effective translation to the end user. This review focuses on the identification of engineering, clinical and user's BMI metrics for new and existing BMI applications.**

I. Introduction

The 2013 International Workshop on Clinical Brain--Neural Machine Interface (BMI) Systems was held on February 25--27, 2013 at the Houston Methodist Research Institute, Houston, Texas [1]-[3]. The purpose of the workshop was to identify and discuss challenges and potential solutions leading to the development and deployment of interface systems based on neural activity in clinical applications. A review of the workshop written by participating trainees can be found in [1].

The challenges identified at the workshop fell into 6 major categories: 1) knowledge gaps in the scientific data regarding long-term device reliability and safety, 2) uncertainty in the regulatory, market and reimbursement pathways, 3) lack of engineering, clinical and patient's metrics for evaluating and quantifying performance in BMI systems, 4) patient-acceptance challenges that impede fast and effective translation to the end user, 5) Lack of established mechanisms for curated data-sharing, and 6) lack of comprehensive clinical, technical and regulatory education and training for the future BMI work force.

In this invited paper, the focus is on the challenge of identifying and defining acceptable BMI metrics for assessing and quantifying performance of new and existing BMI systems. The exposition below summarizes the discussion by participants at the Houston's workshop [1]. Although efforts have been made to provide an impartial and comprehensive review of the spectrum of opinions voiced by the participants at the workshop, the author assumes responsibility for any errors or omissions in this short review.

The identification and selection of suitable metrics to assess BCI performance has been recognized as an important challenge not only to properly evaluate the BMI device but also to allow comparisons between different BMI systems or between similar but non-identical tasks [4]. In this regard, public efforts have been recently made, including the Workshop on BCI metrics at the Asilomar meeting held on June 3-7, 2013 [5], which is summarized in [6]. A recent study have also addressed some challenges and limitations in the development and selection of BCI performance metrics [7], including developing efficient measurement techniques that adapt rapidly and reliably to capture a wide range of performance levels and the identification of BCI subsystems that may potentially restrict the maximum systems level performance, which is a critical factor for considerations of device interoperability. As the definition of metrics for BMI systems is a work in progress, any interested party is encouraged to contact the author to provide comments, suggestions or otherwise get involved in on-going efforts for defining standard metrics for BMI systems. Due to space limitations, the reader is referred to introductory articles on brain-computer interfaces [8], shared control [9], and information transfer rate in BCIs for communication [10].

II. Definition of BMI Metrics

A. Evaluating Patient-Centered Outcomes in BMI Systems

The ultimate goal for all BMI technology is to improve the quality of life and well being of the patient populations who use the technology while reducing the cost of healthcare. Current clinical outcome measures may not reflect the overall benefit that the BMI systems brings to the patient nor they accurately capture the functional gains as interpreted by the patient in a real-world context. Horwitz and colleagues [11] have emphasized that clinical research studies should be designed to more closely approximate real-world use of therapeutics and biomedical devices. They note that in the pursuit of a valid answer, randomized controlled trials "that emphasize efficacy under near-ideal conditions have become a preferred strategy for both regulators (who need to approve medicines and devices for clinical use) and investigators (who design trials). When “efficacy trials” dominate, and studies that reflect real-world use of the treatment are reduced in importance, a surprising collateral effect is that the value attributed to the patient’s experience with their disease and its treatment is diminished" [11].

*Research was supported in part by NIH R01NS075889, IIS-1219321, R01 NS081854, IIS-1302339, IIS-1313620, and R13 NS082045-01.

J. L. Contreras-Vidal is with the Department of Electrical and Computer Engineering, University of Houston, TX 77004, USA and the Department of Neurosurgery at The Methodist Hospital Research Institute (e-mail: jlcontreras-vidal@uh.edu).

978-1-4799-3840-7/14/$31.00 ©2014 IEEE

At the Clinical BMI workshop, participants agreed that different clinical populations such as stroke, ALS, amputees or SCI patients might prioritize differently their needs, challenges, and have different benefit/risk profiles. For example, in terms of accepting a certain degree of invasiveness in the BMI system, or a desired operating speed of the BMI device. Moreover, patients may also evaluate BMI devices in regard to usability (e.g., maintenance requirements of the system, set-up time, cosmesis, etc.), functional gains as well as other psychological factors that influence patient's acceptance of the technology.

Participants at the workshop suggested that existing metrics could be adopted by the BMI community, including: 1) The International Classification of Functioning, Disability and Health (ICF), which is a classification of health and health-related domains that also includes a list of environmental factors to address functioning and disability of an individual that occurs in an environmental context [12]. The ICF is the international standard to describe and measure health and disability. Importantly, metrics have both clinical and regulatory relevance and must address:

a. Determination of the neurological profile of individuals who are capable of using a specific BMI device (including the prosthetic device).
b. Determination of the incidence of adverse effects in the use of the BMI system.
c. Determination of the extent of mobility or function achieved by the use of the BMI system.
d. Determination of any measurable health benefits with the use of the BMI system.
e. Determination of improvement of quality of life with the use of the BMI system [29].

In regard to safety, it is important to note that robotic exoskeletons and other wearable prosthetics may impose unusual joint kinetics and kinematics that could potentially injure bone or skin, particularly in SCI or stroke populations that characteristically have accelerated osteopenia or osteoporosis, unusual spasticity patterns, abnormal movement synergy patterns, or contractures [15]. While impedance control, motion limited to the physiological range of motion and torque cut-offs can greatly mitigate risks and increase safety in upper and lower extremity robotics, cumulative experience is still very limited for mobility devices, warranting caution and careful consideration to appropriately apply this exciting new technology.

2) The System Usability Scale (SUS, [13]), which provides a "quick and dirty", reliable tool for measuring the usability of a wide variety of products and services, including hardware, software, mobile devices, websites and applications. It consists of a 10-item questionnaire with five response options for respondents; from Strongly agree to Strongly disagree. The SUS has become an industry standard, and it is a very easy scale to administer to participants, and it can be used on small sample sizes with valid and reliable results [13].

3) The Technology Readiness Levels (TRL, [14]), which is a type of measurement system used to assess the maturity level of a particular technology. Each technology project is evaluated against the parameters for each technology level and is then assigned a TRL rating based on the projects progress. There are nine technology readiness levels. TRL 1 is the lowest and TRL 9 is the highest [14]. Importantly, the technology development process transitions throughout the life of the project, and a safety strategy input is required early in the project life cycle as part of the technology development process.

B. Metrics for Evaluating Performance in BMI Systems

Ideally, BMI systems should be reliable, effective (i.e., BMI performance should be adequate for the target clinical population), robust, allow for multitasking, require minimal effort, and release attentional resources to other cognitive-motor tasks that the patient may want to get involved in, e.g., speech, eating, etc. Accordingly, engineering metrics for BMI performance should consider all these aspects:

B1. Reliability: The goal is to define metrics that can assess how reliably and robustly a closed-loop BMI can operate a wearable prosthetic. The reliability should be assessed on the complete system (including the patient in the loop), although reliability of system components may also be useful for modular designs. Unfortunately, with a few exceptions, reliability has not been a focus of prior research. Simeral et al [16] and Chadwick et al [17] have examined the stability and reliability of intra-cortical microelectrode array recordings/decodes in a human with tetraplegia 1000 days post-implantation using performance measures in a cursor control task during 5 consecutive days [16] or the control of simulated 2D arm reaching at days 1049, 1057 and 1080 post-implant [17]. Chao et al. examined the robustness of neural representations and signal -to-noise ratios (SNR) of ECoG recordings over a period of months using decoding of hand position and arm joint angles during reaching in non-human primates [18]. They found that decoding did not degrade significantly over this relatively short time, and reported that decoding performance and time between model generation and model testing were not negatively correlated. These studies however do not elucidate the system's reliability and robustness outside the short reporting periods nor inform us of any sources of failures encountered throughout the current lifetime of the implant.

In this regard, physics-of-failure analysis with respect to expected life cycle stresses & lifetime, syndromic monitoring studies, and the design sensor canaries for self-diagnostic of signal quality may be required for characterizing the reliability and robustness of a BMI prosthetic. In addition, methods for real-time anomaly detection and error correction, as well as methodologies for estimating model uncertainty using model performance data are needed. Some metrics that can be deployed are:

Reliability metric: The *operational system availability* of the BMI system, addresses the continued dependence of the patient on the neural interface for the execution of ADLs.

Availability metric: It reflects the probability that the *system* will operate satisfactorily at time *t* when called upon for use. It is expressed as the total system up time divided by the total operating hours. Of course, high reliability and

availability electronics can be achieved on the basis of predicting the possible failure site, failure mode, and mechanism of bioelectronics systems. The detection of first faults during operation for fault resistance and fault tolerance with systems that are capable of monitoring and transmitting degradation related signals is important to determine possible future loss of functionality.

B2. Performance: It is generally agreed that the BMI literature and methods are difficult to evaluate and compare due to differing experimental protocols, evaluation metrics, assumptions, source signals, use of shared control, number of electrodes or features, feedback modality used, length of training, type of decoding (neural classification vs. continuous time trajectory decoding, etc.) and even types of users tested [19]. Typically, engineering metrics for assessing decoding performance in BMI systems have been mostly limited to a few: 1) transfer of information by BMI systems [19]-[20], 2) accuracy (e.g. Pearson's correlation coefficients [21], or the signal to noise ratio (SNR) between the measured and the predicted decoder output using cross-validation techniques [22]), and less often, neural tuning or neural adaptation to BMI use [23].

1) Transfer of information by BMI systems (or information transfer rate, ITR). ITR (bits per sec) is a general evaluation metric devised for brain-computer interface systems (BCIs) for restoring linguistic communication such as P300 BCI spellers (see [19] for a review) or to evaluate performance in BMI systems for 2D cursor control [10], [20]. It also allows comparison of the performance of BCI systems, which have a different number of tasks [6], [7], [21]. Speier et al [19] summarize several limitations of the current use of ITR as a BMI metric: a) conditional probabilities for selection sequences have not been reported, b) information about types of errors in BCI for communication are not used to improve their selection (errors are either ignored or deleted; time outs in 2D BCIs limit quantification of performance), c) task constraints or 'shared control' are usually not factored in the quantification of BMI performance, and d) it is unclear how low the ITR would need to be in order to understand the BCI output. In addition, ITR assumes that there is only *one* information channel that can be used to extract information from the brain, and it is not clear how ITR could be used to quantify performance in a neuroprosthetic limb performing continuous decoding for robot control rather than neural discrete classification of targets.

With respect to BMI for cursor control, Tehovnik et al reviewed the literature and reported, "Typically, the bit rate of the [reviewed] BMI studies fell below 1 bit per second" (page 137, [20]). These studies included human and non-human primate subjects based on single cell, ECoG or EEG sources. Moreover, it was reported that the amount of information transfer with BMI saturates after about 50 neurons when using fixed electrode arrays. The limited performance however could be addressed by taking into account the limitations mentioned above for BCI for communication as they also apply to BCIs for cursor control.

2) Accuracy: While both synchronous and self-paced BMI systems based on discrete classification of neural signals have been evaluated using various metrics (bit rate, confusion matrix, sensitivity and specificity, and others; please see [21] for a detailed review), model-based continuous state decoders inferring continuous time kinematics or kinetics normally use a Kalman or Wiener filter to translate neural activity into motor commands [22], [24]. Performance evaluation is typically done off-line (e.g., during calibration or training of the decoder although some real-time variants have also been proposed that do not differentiate between training and performance) using cross-validation procedures with Pearson's correlation coefficient (*r)*, the coefficient of determination (r^2), or the SNR values used to assess the quality of the reconstructed kinematics or kinetics [22], [24]. One limitation of off-line decoding is the observation that a BMI's prediction power does not necessarily translate into improved closed-loop BMI performance (see [25] for a discussion), and thus metric reporting on off-line decoding performance may not be a suitable metric for BMI systems. In recent closed-loop BMI-cursor control systems, accuracy has been evaluated using the error rate (ER), which measures the percentage of the runs where a target is missed (a target could be missed because either a time limit expired or a false target was selected, [26]). Existing standards could be incorporated in BMI metrics, including ISO 9241, part 9 standard [27] for testing pointing device performance and user assessment as has been recently reported in a BMI system for point-and-click cursor control for humans with tetraplegia [26].

3) Neural tuning or adaptation: It has been noted that use of a BMI system may trigger or enhance neuroplasticity as the user learns to control the neuroprosthetic system or adapts to controlled perturbations within the BMI task environment ([23], for a review see [25]). As brain adaptation is a desirable property of BMI systems for assistive and rehabilitative applications, engineering and clinical metrics capturing neural tuning or adaptation, and relating those to clinically and user meaningful benefits may be useful in comparing closed-loop BMI systems. Metrics that examine how each neuronal unit (or electrode, or region of interest) modulates its firing rate (or neural activity) with respect to discrete and/or continuous states across sessions in BMI longitudinal studies are likely to provide the most useful information [26], [28].

II. Conclusions

The goal of a BMI system is to extract the intent or goals from the user's neural activity and to provide reliable control outputs to external devices leading to quantifiable functional gains. Evaluation of diverse BMI systems will require careful selection of user, clinical and engineering metrics, which could ultimately assist in the BMI development, comparison of BMI systems, and prediction of user acceptance of a given BMI system. In this short review, several metrics have been discussed, however, these metrics should be seen as complementary, and proper analysis of the closed-loop BMI system performance should consider them

as a whole. Moreover, it should be stressed that it is not clear how these different metrics should be weighted in order to compare different alternative solutions, as this is clearly highly dependent on the application. Thus, the evaluation of these systems is inherently multidisciplinary and all relevant stakeholders, including end-users should be involved in it.

REFERENCES

[1] Liew SL, Agashe H, Bhagat N, Paek A, Bulea TC. (2013). A clinical roadmap for brain-neural machine interfaces: trainees? Trainees' perspectives on the 2013 international workshop. *IEEE Pulse*. 4 (5), 44; http://bmiconference.org/

[2] Sukel K (2013). The Dana Foundation;; 'Roadmapping the Adoption of Brain--Machine Interfaces', April 15, 2013; http://www.dana.org/news/features/detail.aspx?id=42098

[3] Neurotech Business Report; 'Brain-Machine Interface Pioneers Participate in Workshop on Clinical BMIs'; http://www.neurotechreports.com/pages/workshop_clinical_brain_machine_interfaces_report.htm

[4] Hill NJ, Häuser AK, Schalk G. (2014). A general method for assessing brain-computer interface performance and its limitations. J Neural Eng. 2014 Apr;11(2):026018.

[5] http://bcimeeting.org/2013/workshops.html; accessed 6/4/2014.

[6] Thompson DE1, Quitadamo LR, Mainardi L, Laghari KU, Gao S, Kindermans PJ, Simeral JD, Fazel-Rezai R, Matteucci M, Falk TH, Bianchi L, Chestek CA, Huggins JE. Performance measurement for brain-computer or brain-machine interfaces: a tutorial. J Neural Eng. 2014 Jun;11(3):035001.

[7] Hill NJ, Häuser AK, Schalk G. A general method for assessing brain-computer interface performance and its limitations. J Neural Eng. 2014 Apr;11(2):026018.

[8] Shih JJ, Krusienski DJ, Wolpaw JR. (2012). Brain-computer interfaces in medicine. Mayo Clin Proc. 2012 Mar;87(3):268-79.

[9] Tonin, L. ; Leeb, R. ; Tavella, M. ; Perdikis, S. ; del Millan, J.R. The role of shared-control in BCI-based telepresence. Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on, pp 1462-1466.

[10] Schlogl, A. ; Keinrath, C. ; Scherer, R. ; Furtscheller, P. Information transfer of an EEG-based brain computer interface. Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on, pp. 641-644.

[11] Horwitz RI, Abell JE, Christian JB, Wivel AE. (2014). Right answers, wrong questions in clinical research. Sci Transl Med. 2014 Jan 29;6(221):221fs5.

[12] http://www.who.int/classifications/icf/en/

[13] http://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

[14] http://www.nasa.gov/content/technology-readiness-level/#.U0_oza1dUcw

[15] Contreras-Vidal JL, Grossman RG. (2013). NeuroRex: A clinical neural interface roadmap for EEG-based brain machine interfaces to a lower body robotic exoskeleton. Conf Proc IEEE Eng Med Biol Soc. 2013 Jul;2013:1579-82.

[16] Simeral JD, Kim SP, Black MJ, Donoghue JP, Hochberg LR. Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array. J Neural Eng. 2011 Apr;8(2):025027.

[17] Chadwick EK, Blana D, Simeral JD, Lambrecht J, Kim SP, Cornwell AS, Taylor DM, Hochberg LR, Donoghue JP, Kirsch RF. Continuous neuronal ensemble control of simulated arm reaching by a human with tetraplegia. J Neural Eng. 2011 Jun;8(3):034003.

[18] Chao ZC1, Nagasaka Y, Fujii N. Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys. Front Neuroeng. 2010 Mar 30;3:3.

[19] Speier W, Arnold C, Pouratian N. (2013) Evaluating true BCI communication rate through mutual information and language models. PLoS One. 2013 Oct 22;8(10):e78432.

[20] Tehovnik EJ1, Woods LC, Slocum WM. (2013). Transfer of information by BMI. Neuroscience. 2013 Dec 26;255:134-46. doi: 10.1016/j.neuroscience.2013.10.003.

[21] Thomas E1, Dyson M, Clerc M. An analysis of performance evaluation for motor-imagery based BCI. J Neural Eng. 2013 Jun;10(3):031001.

[22] Fitzsimmons NA, Lebedev MA, Peikon ID, Nicolelis MA. Extracting kinematic parameters for monkey bipedal walking from cortical neuronal ensemble activity. Front Integr Neurosci 3: 3, 2009.

[23] Ganguly K, Carmena JM. Emergence of a stable cortical map for neuroprosthetic control. PLoS Biol. 2009 Jul;7(7):e1000153.

[24] Presacco A, Forrester L, and Contreras-Vidal JL. Decoding Intra-Limb and Inter-Limb Kinematics During Treadmill Walking From Scalp Electroencephalographic (EEG) Signals, IEEE Trans Neural Syst Rehabil Eng, 20(2): 212–219, 2012.

[25] Orsborn AL, Carmena JM. Creating new functional circuits for action via brain-machine interfaces. Front Comput Neurosci. 2013 Nov 5;7:157.

[26] Sung-Phil Kim, John D. Simeral, Leigh R. Hochberg, John P. Donoghue, Gerhard M. Friehs, and Michael J. Black. Point-and-Click Cursor Control With an Intracortical Neural Interface System by Humans With Tetraplegia IEEE TNSRE, VOL. 19, NO. 2, APRIL 2011, 193-203

[27] S. A. Douglas, A. E. Kirkpatrick, and S. I. MacKenzie, "Testing pointing device performance and user assessment with the ISO 9241, part 9 standard," in Proc. ACMSIGCHI Conf. Human Factors Comput. Syst. (CHI'99), 1999, pp. 215–222.

[28] Kilicarslan, Atilla; Prasad, Saurabh; Grossman, Robert G.; Contreras-Vidal, Jose L., "High accuracy decoding of user intentions using EEG to control a lower-body exoskeleton," 35th Annual International Conference of the IEEE, vol., no., pp.5606,5609, 3-7 July 2013.

[29] J.H. Ware, C. Sherbourne. The MOS 36-Item Short-Form Health Survey (SF-36). Medical Care 30: 473-83, 1992.