THE EUROPEAN SOCIETY OF MEDICINE
Medical Research Archives, Volume 13 Issue 1

RESEARCH ARTICLE

Biomarkers in clinical practice: opportunities and challenges

Paul E. Rapp, Ph.D.[1*], Adele M. K. Gilpin, Ph.D., J.D.[2]

[1]Department of Military and Emergency Medicine, Uniformed Services University
[2]Gilpin Phillips BIOMED, LLC

*paul.rapp@usuhs.edu

OPEN ACCESS

PUBLISHED
31 January 2025

CITATION
Rapp, PE. and Gilpin, AMK., 2025. Biomarkers in clinical practice: opportunities and challenges. Medical Research Archives, [online] 13(1). https://doi.org/10.18103/mra.v13i1.6251

DOI
https://doi.org/10.18103/mra.v13i1.6251

ISSN
2375-1924

ABSTRACT

Many proposed biomarkers fail to produce clinically actionable results. Simply put, the research problem addressed here is: why do most biomarker projects fail? In this contribution we describe four commonly encountered problems and outline procedures that address these challenges. The specific issues addressed are as follows:

1. A statistically significant result in a between-group hypothesis test often does not result in classification success.
2. Cross-validation is commonly used in model validation. The successive steps in cross-validation expose it to multiple sources of failure that may result in erroneous conclusions of success.
3. Failure to rigorously establish the test-retest reliability of a biomarker panel precludes its use in longitudinal monitoring of treatment response or disease progression. Further, it should be recognized that the minimum detectable difference is not the minimal clinically important difference.
4. Sample size estimates used in the design of a clinical study must be determined by the objectives of the study. The sample sizes required in reliability studies and in the evaluation of biomarkers as prodromes must be determined with those objectives in mind and are far larger than sample size requirements computed for the purpose of hypothesis testing.

We conclude with suggestions for transparency and collaboration that would facilitate the use of biomarkers in clinical practice.

I. Introduction

The use of biomarkers in clinical practice presents extraordinary opportunities but also introduces significant challenges. In this contribution we describe several issues that must be addressed in the biomarker discovery process. What follows is not a comprehensive review of the literature but rather is based on clinical observation with supporting evidence from the referred literature. In the current literature the term "biomarker" is often used casually. This is changing. In the United States the FDA-NIH (Food and Drug Administration – National Institutes of Health) Biomarker Working Group[1] provided a definition of a biomarker "A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention". This definition is expanded in this document to include definitions of biomarker types: diagnostic, monitoring, pharmacodynamic/response (includes surrogate endpoints), predictive, prognostic, safety, and susceptibility/risk biomarkers. An FDA pathway for biomarker qualification has been established.[2] When more demanding criteria are introduced, the number of successes falls. For example, Prata, et al.[3] examined 3221 papers investigating psychosis-related biomarkers. Only one biomarker passed their criteria for clinical applicability (an SNP on HLA-DBQ1 predicted clozapine-induced agranulocytosis).

Broadly stated biomarker utilization is a modeling and statistical process. A large number of measures are obtained from, hopefully, well defined populations (for example, generalized anxiety disorder, bipolar disorder, healthy comparison groups, etc.). The same measures are obtained from the patient of interest and the operational question is what is the membership probability of this individual in each of the reference groups? In practice biomarkers have at least four objectives:

(i.) diagnosis

(ii.) longitudinal measurement of treatment response or disease progression,

(iii.) identification of individuals at risk of disease onset (prodrome discovery), and

(iv.) matching specific patients to an appropriate treatment

On reflection it is seen that all of these objectives are classification problems. Diagnosis is clearly a classification problem. Longitudinal assessment can be treated as a classification problem by asking, is the probability of an individual's membership in an appropriately constructed healthy comparison population increasing or decreasing with time? Similarly, the identification of individuals at risk of disease onset can be treated as a classification problem by asking what is the probability that this patient is a member of a population of individuals that progressed from a sub-clinical state to a clinical presentation? Matching a specific patient to a course of treatment can be conceptualized as a special case of diagnosis. In this case, the operational question is, what is the probability that this individual is a member of a population that responded positively to this treatment option.

Concerning classification problems, two observations immediately present themselves. First, several different mathematical classification methods are available, and there is no single best method. Attention is directed to the study by Fernández-Delgado et al.[4] with the challenging title "Do we need hundreds of classifiers to solve real world problems?" They compared 179 classifiers against 121 data sets in the University of California, Irvine database. They make specific recommendations. We also note that best practice is to repeat all classification problems with different algorithms. Second, as will be argued, the size of the clinical and comparison reference populations required for statistically responsible classification can be very large. This requirement may be a critical unresolvable problem for biomarker utilization.

II. Diagnosis

The definitional problems associated with diagnosis, particularly in neuropsychiatry, makes it the most challenging of the four biomarker objectives.

Nonetheless, this is often where biomarker research begins. Guidance for reporting diagnostic accuracy studies are available[5] and, contrary to frequent practice, should receive systematic application in biomarker investigations. At a minimum, the evaluation of a candidate biomarker should not be limited to sensitivity and specificity, but should be expanded to include positive and negative likelihood rates, positive and negative predictive rates, false discovery rates, and the area under the ROC (receiver operating characteristic) curve. These reports must include confidence intervals.

Willful or inadvertent errors in diagnostic hypothesis testing frequently occur. Head, et al.[6] concluded that "p-hacking is widespread in science." In part this reflects a misunderstanding of the p-value. A growing body of opinion maintains that better alternatives exist. This is considered in the American Statistical Association's statement on p-values[7]. A vigorous summary of the views of some statisticians is given in Wasserstein, et al.'s "Moving to a world beyond p-values"[8]. Additional opinions should also be considered, however[9]. Extensive documentation addressing the discussion is available online at the American Statistical Association.

In the specific context of classification, the important assessment is probability of classification error, P_{ERROR}. A low between group p-value does not ensure successful classification. Rapp, et al.[10] constructed a two-group classification example where $p = 2 \times 10^{-11}$, but $P_{ERROR} = 0.4078$. Recalling that the error rate of random classification in a two-group classification is $P_{ERROR} = 0.5$, it is seen that little better than random performance was achieved. In the case where a single continuous variable is used in a two-group classification, and where the distributions of the discriminating variable in each group are normal (or near-normal) it is possible to predict P_{ERROR} with the number of participants in each group, and the group means and standard deviations. This makes possible a re-assessment of some previously published candidate biomarker papers. These post-facto calculations can be disheartening.

III. Model selection and validation

Biomarker classifier performance can often be improved by incorporating more variables into a multivariate classification, but more is not necessarily better. Watanabe, et al.[11] published an EEG classification example where P_{ERROR} decreased as measures were eliminated from the classifier.

Mathematically informed model selection is essential. Model selection is a process in which candidate measures are either incorporated into or rejected from a classifier. LASSO (Least Absolute Shrinkage and Selection Operator)[12] is often used in biomedical research and has the advantage of preventing overfitting and presenting results in a form that is readily interpretable. An extension of this method, elastic net model selection[13] can outperform LASSO on some problems (Zou and Hastie[13] show that LASSO is a special case of elastic net LARS-EN). Random forest methods can also be used in model selection[14,15]. As in the case of classification algorithms, the recommendation is to perform biomarker model selection with more than one algorithm. Significant divergences in selection should be investigated.

Once a model has been constructed, cross validation is a commonly used procedure for model validation. This procedure is not robust against misapplication. Indeed, errors in cross validation are so common that the standard textbook for statistical learning, Hastie, et al.,[16] includes section 7.10.7 "The wrong and the right way to do cross validation." Misapplied, cross validation can produce sensitivity, specificity and positive prediction values in excess of 0.95 with random numbers.

IV. Longitudinal monitoring, quantifying change

A biological measure becomes a candidate biomarker if its values depart from normal values in a consistent manner when obtained from a well-defined clinical population. Ideally, a candidate monitoring biomarker would accurately track clinically perceived patient status. For example, the value of the biomarker would increase as the patient improved. Reality is not so kind. In some instances, the value of the candidate biomarker is largely unchanged even through the patient is improving. Kemp et al.[16] provides an example. Heart rate variability was found to be reduced in depressed patients but did not change when patients improved clinically in response to medication. There are several reasons that can explain the failure of a candidate monitoring biomarker to track clinical status. The first, and possibly most common, is that the original identification as a biomarker was flawed. Sample sizes in clinical studies can result in false positive identification of biomarkers. Hajcak, et al.[17] have directed attention to another possibility. It is possible that the candidate biomarker is not correlated with the disease process itself but rather is correlated with a risk factor for the disease. When this occurs, populations studies will identify a correlation between the presence of the biomarker and the disease, but the biomarker will not change in response to treatment.

It is commonly argued that even if a biomarker's between-diagnostic-group specificity is low, it may still be useful when used longitudinally to monitor patient status. Body temperature is an example. Used alone temperature has low diagnostic specificity, but used longitudinally can be an important indication of improvement or deterioration. Temperature is an effective monitor because (1.) thermometers give precise measurement of temperature, and (2.) temperature is stable in a clinically stable individual. Thus, a critical step in establishing the utility of a candidate monitoring biomarker is determination of its test-retest reliability. In the case of psychophysiological biomarkers, the literature on test-retest reliability is limited and discouraging. This can also be true of neuropsychological assessments. For example, Cole, et al.[18] conducted a reliability study of four neuropsychological batteries used in pre- and post-concussion assessments. They concluded, "However, small test-retest reliabilities in four NCATs (Neuropsychological Assessments Tests) in a military population are consistent with reliability reported in the literature (non-military populations) and are lower than desired for clinical decision making."

Reliability studies of biomarkers are now an urgent requirement. Linear correlation should not be used to quantify reliability. Reliability for continuous data should be quantified by the intraclass correlation coefficient, ICC. It is important to note that there is more than one variant of the ICC. Shrout and Fleiss[19] have identified six versions, and McGraw and Wong[20] give ten versions. The choice depends on the details of the evaluation protocol. Müller and Büttner[21] and Koo and Li[22] provide directions for selecting among the possible ICC versions. In a report of ICC results it is essential to identify the version used and to include confidence intervals.

When the intraclass correlation coefficient is available, change in a biomarker can, to a degree, be interpreted by calculating the corresponding minimum detectable difference, MDD, and the standard error of measurement, SEM. It must be understood that the MDD and the SEM characterize properties of measurement distribution. They are not equivalent to the minimal clinically important difference that has to be calculated by very different procedures[23]. Reliability as quantified by the ICC is necessary but not sufficient for clinical utility. Interpretation of the ICC is further complicated when it is recognized that the ICC is specific to a population. Young healthy controls typically have the greatest reliability. Some clinical populations have low reliability, for example head injury patients[24]. A change in a biomarker that is unremarkable in an elderly patient could potentially be a matter of concern in a young adult.

V. Sample sizes in biomarker validation

Inadequately powered studies are a severe and continuing issue in neuroscience. In "Power failure: why small sample size undermines the reliability of neuroscience," Button, et al.[25] reported, "Here we show that the average statistical power of studies in neuroscience is low. The consequence of this includes overestimates of effect size and low reproducibility." We have published an example from our own research group[26]. The investigation was directed to determining if event related potentials obtained from returning service personnel could identify those who would present delayed onset PTSD in six months or one year after return from combat. The initial analysis indicated Sensitivity=0.8 and Specificity=0.87. Prior to publication, a further analysis confirmed those values of sensitivity and specificity but found that the corresponding confidence intervals were [0,1]. The sensitivity and specificity values were a fortuitous consequence of a small sample size. This example emphasizes two points; small sample sizes can produce false positive indications of biomarker utility, and, second, the determination of confidence intervals is an essential element of every analysis.

Sample size requirements are specific to the objective of the investigation. Zou[27] has provided guidance for intraclass correlation coefficient test-retest studies. It should be noted, and Professor Zou has acknowledged, that there are a number of typographic errors in the equations in this paper. Our calculations have independently confirmed the contents of the sample size tables in Zou. Notably, the sample sizes identified by this analysis are much larger than those typically reported in biomarker test-retest studies.

Prodrome studies are constructed by identifying a population of individuals who do not meet diagnostic threshold for a disorder but where there is a reason to believe that a significant fraction of that population will subsequently present the disorder (an enriched population). The sample size required for this study will be critically dependent on the fraction of the intake population that will present the disorder (the converter versus sable fraction). If the converter fraction is low, which is often the case in neuropsychiatric disorders, the sample size required can be large. Consider a conversion rate of 10%, which is higher than typically obtained, and an acceptable uncertainty in sensitivity of ± 0.1. Hoeffding's inequality[28] requires 185 converters and therefore a study population of 1850. Hoeffding's criterion fixes the precision of the interval independently of the value of sensitivity and therefore gives a conservative (greater) estimate of the required sample size than the criterion constructed by Clopper and Pearson[29] which requires 104 converters and a total population of 1048. The requirement for large sample sizes in prodrome studies, which by definition are longitudinal studies and therefore expensive, may be a critical limiting factor in prodrome discovery.

VI. Paths forward

The challenge of mathematical analyses of data obtained in biomarker studies should not be underestimated. An instructive example has been published by Botvinik-Nezer, et al.[30]. In this study the same fMRI data set was sent to 70 independent teams for analysis. They report that "no two teams chose identical workflows to analyze the data. This flexibility resulted in sizable variations of the results of hypothesis tests, even for teams where statistical maps were highly correlated at intermediate stages of the analysis pipeline." No suggestion is made of incompetence or misbehavior. Simply put, the challenges of large data sets be it imaging data, electrophysiological signals, genetic studies or their combination are formidable.

In the specific case of psychophysiological biomarkers, the Society for Psychophysiological research made several recommendations that readily generalize to other types of biomarkers. The recommendations included data sharing with data format harmonization, analysis software sharing and preregistration of analysis plans[31,32]. The free availability of powerful analysis software carries the danger of misapplication. To a degree, this can be addressed by applying software to be used in an anticipated study to publicly available data as part of the software validation process. Public data sources are expanding daily and include ERP data[33], connectome data[34], MEG data[35], EEG data [36], and genomic data (the NIH GenBank).

It has been proposed that preregistration of a research analysis plan can reduce bias, increase transparency and facilitate reproducibility of research[32,37]. Registration of an analysis plan is not limited to a description of computational procedures. It includes description of the study design, motivating hypotheses, variables, and data collection methods as well as planned statistical analysis. This document can be submitted to a public online repository such as the Open Science Framework. Preregistration resources are available at the American Psychological Association website.

Systematic implementation of data sharing, software sharing and preregistration will not come easily or quickly to the biomedical research community. We nonetheless suggest that the path forward in biomarker discovery will not only require scientific advances but also cultural change.

VII. Conclusion

An historical opportunity to advance clinical practice with biomarkers is now available to us. Data acquisition capabilities have expanded exponentially. Computational resources have expanded at a similar rate. Additionally, the essential underlying mathematical methods needed for large scale data analysis have been constructed and validated. The clinical realization of this opportunity has, however, expanded at a slower pace. Many factors contribute to this. In this contribution we identified four areas that have the advantage of being immediately addressable.

1. Appropriate mathematical methods for classifying individuals into recognized clinical populations have been developed.
2. It is recognized that the mathematical methods of statistical learning are not robust against misapplication, but procedures to prevent misapplication have been identified and can be implemented.
3. The critical need for systematic test-retest reliability studies has been established, and appropriate procedures for quantifying reliability are available.
4. Mathematical results establishing the need for much larger sample sizes in biomarker studies have been established, and sample size guidelines specifically for biomarker studies have been published.

The magnitude of the challenges ahead should not be underestimated, but, conversely, the magnitude of the opportunity should not be underestimated.

Author Contributions:

P.E.R. conceptualization, literature review, first draft; A.M.K.G. revision of the final draft. Both authors have read and agreed to the published version of this manuscript

Conflicts of Interest:

The authors declare no conflict of interest.

Funding:

This research did not receive external funding

Institutional Review Board Statement:

Not applicable. This contribution is a review of publicly accessible previously published literature.

Informed Consent Statement:

Not applicable

Data Availability Statement:

Not applicable

Disclaimer:

The opinions and assertions contained herein are those of the authors and do not necessarily reflect the official policy or position of the Uniformed Services University, the US Department of Defense, or the Henry M. Jackson Foundation for the Advancement of Military Medicine.

References:

1. Food and Drug Administration, FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS and Other Tools) resource. 2021; Silver Spring: Food and Drug Administration.

2. Food and Drug Administration. Biomarker qualification: evidentiary framework. Guidance for Industry and FDA staff. Draft Guidance FDA 2018.

3. Prata D Mochelli and Kapur S Clinically meaningful biomarkers for psychosis: a systematic and quantitative review. *Neurosci and Biobehavioral Rev*. 2014; 45, 134-141.

4. Fernández-Delgado M, Cernadas E. and Barro S. Do we need hundreds of classifiers to solve real world classification problems? *J. of Machine Learning Res*. 2014,15, 3133-3181.

5. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatson CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016; 6: e012799.

6. Head ML, Holman L, Lanfear R, Kahn AT and Jennions MD The extent and consequences of p-hacking in science. *PLoS Biology*. 2015; 13(3): e1002106.

7. American Statistical Association. The ASA's statement on p-values: context, process and purpose. *Amer. Statistician*. 2016; 70(2), 129-131.

8. Wasserstein RL, Schirm AL, and Lazar NA Moving to a world beyond "p<0.05." *Am Statistician*. 2019; 73, Supplement 1, pp 1-19. Article 1.

9. Ionides, EIL, Giessing, A., Ritov, Y and Page, SE. Response to the ASA's statement on p-values context, process and purpose. *The Amer. Statistician*. 2017, 71(1), 88-89.

10. Rapp PE, Cellucci CJ, Keyser DO, Gilpin AMK and Darmon DM Statistical issues in TBI clinical studies. *Front in Neurology*. 2013; 4, 177.

11. Watanabe TAA, Cellucci CJ, Kohegyi E, Bashore TR, Josiassen RC, Greenbaun NN and Rapp PE The algorithmic complexity of multichannel EEGs is sensitive to changes in behavior. *Psychophysiology*, 2003 40, 77-97.

12. Tibshirani R Regression shrinkage and selection via the Lasso. *J Roy Stat Soc. Series B*. 1996; 58(1), 267-288.

13. Zou H and Hastie T Regularization and variable selection via the elastic net. *J Roy Stat Soc Series B*. 2005; 67(2), 301-320.

14. Fox EW, Hill RA, Leibowitz SG, Olsen AR, Thornburg D.J. and Weber MH (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environ Modeling Assess*. 2017; 189(7), 316.

15. Speiser JL, Miller ME, Tooze J. and Ip E A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems Applications*. 2017; 1234, 93-101.

16. Hastie T, Tibshirani R and Friedman J Elements of Statistical Learning. Second Edition. New York: Springer. 2009

17. Hajcak G. Klawohn J and Meyer A The utility of event-related potentials in clinical psychology. *Ann Rev Clin Psych*. 2019; 15, 71-95.

18. Cole WR, Arrieux JP, Schwab K, Ivins BJ, Qashu FM and Lewis S. Test-retest reliability of four computerized neurocognitive assessment tools in an active duty military population. *Arch of Clin Neuropsych*. 2013; 28, 732-742.

19. Shrout PE and Fleiss JL Intraclass correlations: Uses in assessing rater reliability. *Psych Bull*. 1979; 86(2), 420-428.

20. McGraw KO and Wong SP Forming inferences about some intraclass correlation coefficients. *Psych Methods*. 1996; 1(1), 30-46.

21. Müller R and Büttner PA critical discussion of intraclass correlation coefficients. *Stat in Medicine*. 1994; 13(23-24), 2465-2476.

22. Koo TK and Li MY A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropractic Res*. 2016; 15(2), 155-163.

23. Copay AG, Subach BR, Glassman S., Polly D. and Shuler T. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J*. 2007; 7, 541-546.

24. Bleiberg J Garmoe WS., Halpern EL Reeves DL and Nadler JD. Consistency of within-day and across-day performance after mild brain injury. *Neuropsychiatry, Neuropsychology and Behavioral Neurology*. 1997,10(4), 247-253.

25. Button KS, Ioannidis JPA, Mokrysz G, Nosek BA, Flint J, Robinson ESJ. and Munafò MR Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neurosci*. 2013; 14(5), 365-376.

26. Wang C, Costanzo ME, Rapp PE, Darmon D, Bashirelahi K, Nathan DE, Cellucci CJ, Roy MJ and Keyser DO Identifying electrophysiological prodromes of post-traumatic stress disorder: results form a pilot study. *Front Psychiat*. 2017; Volume 8, Article 71.

27. Zou GY Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Stat in Medicine*. 2012; 31, 3972-3981.

28. Hoeffding W Probability inequalities for sums of bounded random variables. *J Amer Stat Assoc*. 1963; 58(301), 13-30.

29. Clopper CJ and Pearson ES The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934; 26(4), 404-413.

30. Botvinik-Nezer B, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. 2020; 582, 84-88.

31. Garrett-Ruffin, S, Cowden Hindash, A, Kaczkurkin, AN, Mears, RP, et al.. Open science in psychophysiology: an overview of challenges and emerging solutions. *Internat J Psychophysiol* 162, 69-78.

32. Hardwicke TE and Wagenmakers EJ Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behav*. 2023; 7(1), 15-26.

33. Kappenman ES, Farrens JL, Zhang W, Stewart AX and Luck SJ ERP CORE: An open resource for human event-related potential research. *Neuroimage*. 2021; 225: 117465.

34. Human Connectome Project. Reference Manual: WU-Minn HCP 500 Subjects +MEG2 Release: WU-Minn Consortium Human Connectome Project. 2014

35. Niso G, Rogers C, Moreauy JT, Chen, L-Y. Madjar C, Das S, et al. OMEGA: the open MEG archive. *Neuroimage*. 2016; 124(pt, B), 1182-1187

36. Van Dijk H, van Wingen G, Denys D, Olbrich S, van Ruth R and Arns M The two decades brain clinics research archive for insights in neurophysiology (TDBRAIN) database. *Scientific Data*. 2022; 9: 33.

37. Nosek BA, Ebersole CR, DeHaven AC and Mellor DT The preregistration revolution. *Proc Nat Acad Sciences*. 2018; 115(11), 2600-2606.