

COMMENTARY

In Pursuit of Greater Reproducibility and Credibility of Early Clinical Biomarker Research

LM McShane

INTRODUCTION

Biomarkers underlie many clinical tests that are integral to the practice of personalized medicine. Reproducibility and scientific credibility of clinical biomarker early development studies are critical to avoid advancing worthless or potentially harmful biomarker-based tests into late-phase clinical studies and clinical practice. This commentary discusses key aspects to consider when conducting and evaluating early clinical biomarker research. Greater attention to these aspects would enhance research reproducibility and better prioritize biomarkers for further clinical development.

Recognition of the problem of irreproducibility of preclinical drug development research led to a call for transparent reporting standards and recommendations for improved study designs.¹ Similar principles apply to early research aiming to develop clinical biomarker tests (henceforth termed “early clinical biomarker research”), but there are important differences too. A major difference between preclinical drug development studies and early clinical biomarker development studies is that the latter are often conducted retrospectively using stored specimens collected in routine clinical care settings or in the context of research studies originally addressing different questions. Thus, early clinical biomarker research has features of retrospective observational studies that extend beyond the experimentally controlled settings typical for preclinical drug development research.

The development process for biomarker-based tests usually begins with a study aiming to establish whether a biomarker is associated with some clinical outcome or other phenotype. The test may be based on a single biomarker or a panel of biomarkers combined via a statistical prediction model; for example, using “omics” assay technologies that measure “related sets of biological molecules in a comprehensive fashion.”² Further development requires a series of studies to gather more evidence and eventually incorporate the biomarker into a clinical test that is validated for a specific clinical use. The clinical role for a biomarker-based test typically falls into one or more of the following categories (see US Food and Drug Administration / National Institutes of Health glossary at <https://www.ncbi.nlm.nih.gov/books/NBK326791/>): diagnostic, monitoring, pharmacodynamics/response, predictive, prognostic, safety, and susceptibility/risk. This commentary focuses on overarching principles to consider in

early clinical biomarker research to enhance reproducibility and provide a solid foundation for later stages of development. For more extensive discussion of best practices to be applied throughout the clinical biomarker development process, readers are referred elsewhere.^{2–4}

Study design and primary data generation

Study design is usually discussed in the context of prospectively conducted preclinical experiments or clinical trials, but many design principles apply also for studies with retrospective elements. Attention should focus on biomarkers that have potential to provide insights into biological processes or translate into tools for clinical decision-making. Building on that foundation, good study design requires recognition of the many factors that can lead to variation in results (systematic and random).

Biomarker assay methods and subject and specimen factors may systematically affect biomarker measurements and their associations with clinical outcomes. In early development studies, biomarker assays should meet at least minimal analytical performance standards to establish that the assay measures the intended analyte and has acceptable reproducibility over the range of values relevant to the clinical setting. Performance criteria become more stringent as the development proceeds (see **Supplementary Table S1**). Assay methods should be documented carefully to facilitate replication. Subject factors such as age or gender; disease status, subtype or stage; and comorbidities should be considered in formulating retrospective eligibility criteria. Requirements for specimen collection, processing, and handling to ensure reliable assay performance should be defined.

Designs confounding important subject- or specimen-related factors with biomarker or outcome status must be diligently avoided. For example, women would be inappropriate control subjects in a study of prostate cancer detection biomarkers. Multi-institutional studies increase the risk of biases. Differences in patient characteristics, clinical management, and specimen handling across institutions may confound associations between biomarker measurements and outcomes. Preferably, studies are designed to avoid such confounding, or at minimum, information should be collected to attempt adjustment for these factors in analyses. Many poorly designed studies exist in the published literature, and their reproducibility is often compromised.

Random factors are those that differ from study to study and generally cannot be controlled completely. Examples

include laboratory assay batches and observers recording subjective biomarker or outcome assessments. “Omics” assays generating large numbers of measurements per specimen are particularly prone to batch effects due to their sensitivity to subtle changes in laboratory conditions.⁵ Effects of random (or not easily standardized) factors can be reduced through randomization and blinding. An example of poor study design is running samples from subjects with a favorable disease outcome in one assay batch and samples from subjects with an unfavorable outcome in another. Samples should be randomized to assay batches or allocated in a way that batch effects could be eliminated through statistical corrections. Observers recording subjective clinical outcomes should not be confounded with biomarker status and should remain blinded to biomarker values. Similarly, insidious biases can occur when individuals making subjective biomarker assessments are not blinded to subjects’ clinical outcomes. Inattention to these design issues can impair study reproducibility.

Sample size (number of study subjects) is another important study design consideration. It may be based on calculated power for a statistical test or precision for an estimate of a parameter of interest. Example parameters are accuracy of a biomarker in identifying individuals who respond to (or experience toxicity from) a drug, or a hazard ratio representing a biomarker’s prognostic association with clinical outcome. Such calculations help to set expectations for evidence to be gained but should be performed prior to study initiation and based on realistic assumptions.

Related to sample size is within-subject replication (number and types of replicate measurements per subject). When measurement error for outcome or other variables is substantial, replication can reduce noise. Example replicate types include biomarker measurements on samples collected over several timepoints and repeated measurements on a single sample. Measurement replication schemes may be important to mimic when attempting to reproduce study results. Total number of observations must not be confused with number of independent subjects, and data analyses must account appropriately for within-subject replication schemes.

Data collection and curation

There are rigorous quality standards for collection and curation of clinical trial data. In contrast, many early biomarker studies rely on clinical characteristics and outcome data collected retrospectively, possibly extracted from clinical charts, registries, or electronic medical record systems. Investigators should make efforts to confirm the validity of such retrospectively collected data to ensure that they are accurate and correctly interpreted. Data from these sources together with newly generated biomarker data also need to be managed with care. Risk of inadvertent data corruption is increased with inexperienced or careless use of software with sorting, cut-and-paste, and autocorrect features. Omics data present additional challenges due to their sheer volume and specialized formats, which require complex data systems managed by experienced personnel.

Data analysis

Many early biomarker investigations are conducted without a statistical analysis plan prespecifying primary analyses or details of analysis approaches. The number of analyses can easily reach dozens considering different end points, subgroups, explanatory variables and models, or cut-points applied to continuous biomarker values. Chances of false-positive findings increase, as each additional analysis may generate false-positive findings from noise in the data. Pitfalls of conducting numerous exploratory analyses are well recognized by clinical trial methodologists.⁶ Outlining key analyses in a prespecified analysis plan helps to distinguish preplanned analyses from data-driven exploratory or *ad hoc* analyses that are more likely to generate false-positive or biased results.

Data analysis approaches must be consistent with study design, including accounting for nonrandom selection of study subjects. Use of case-control and matched study designs is fairly common in retrospective biomarker studies, and these require specialized statistical analysis methods.^{7,8} Analyses should additionally account for multiple testing, data distributions (e.g., nonnormal data), functional relationships between biomarkers and outcomes (e.g., nonlinear), correlations between multiple measurements per study

Table 1 Reporting guidelines particularly relevant to clinical biomarker research

Acronym	Reporting guideline title	Website	Study type
BRISQ	Biospecimen reporting for improved study quality	http://www.equator-network.org/reporting-guidelines/brisq/	Studies utilizing biospecimens
CONSORT	Consolidated standards of reporting trials	http://www.consort-statement.org/	Randomized clinical trials
REMARK	Reporting recommendations for tumor marker prognostic studies	http://www.equator-network.org/reporting-guidelines/reporting-recommendations-for-tumour-marker-prognostic-studies-remark/	Tumor marker prognostic studies (and prognostic studies more generally)
STARD	Standards for the reporting of diagnostic accuracy studies	http://www.equator-network.org/reporting-guidelines/stard/	Diagnostic accuracy studies
STROBE	Strengthening the reporting of observational studies in epidemiology	http://www.equator-network.org/reporting-guidelines/strobe/	Observational studies in epidemiology (and more generally)

subject, and handling of outliers and missing data. Statistical analyses cannot rescue data that are corrupted or generated by terribly flawed study designs; in the opposite direction, inappropriate statistical analyses can lead to misleading results and inappropriate conclusions even when based on high-quality data.

Results interpretation and study reporting

Complete and transparent reporting of study design, conduct, analysis, and results facilitates proper interpretation of a study and evaluation of its quality. Others may be unable to reproduce results of a study if not adequately informed about the study population, specimen requirements, and biomarker assay methodology. Different data analysis approaches may lead to different results, so it is important to describe analyses that were performed and why those approaches were selected. Disclosure of the total number of analyses performed and which were prespecified is important to gauge potential for false-positive findings. Study sample size and precision of estimated effects or parameters of interest should be reported to indicate the strength of evidence; for example, to help distinguish nonsignificant from convincingly null findings. Relevant parameters to report will differ depending on the potential clinical role for the biomarker. For example, a metric reflecting discrimination ability or accuracy is more relevant than one reflecting association for a candidate diagnostic biomarker.

Detailed guidance for reporting a variety of types of health research studies is available on the EQUATOR website (<http://www.equator-network.org/reporting-guidelines/>). Several of particular relevance to biomarker studies are listed in **Table 1**. Although reporting guidelines do not dictate how research should be performed, many investigators find them helpful to consult when planning studies to be reminded of critical aspects of study design, conduct, and analysis to consider.

Results dissemination

The tendency to preferentially publish studies showing positive or statistically significant findings is known as publication bias. A related phenomenon is selective reporting of results within a study (e.g., only for certain outcome measures or subgroups among many examined), where usually those reported are statistically significant, especially in a desired or expected direction. Evidence for publication bias and selective reporting in clinical trials has been firmly established.⁹ For early clinical biomarker research, the potential for biases is greater due to lack of an organized system for study registration (analogous to ClinicalTrials.gov for clinical trials) and typical absence of comprehensive study protocols with prespecified statistical analysis plans. For every biomarker study reporting positive results, it is unknown how many studies of the same biomarker failing to achieve desired or statistically

significant results never saw the light of day, or what resources were expended on failed or unreported studies. Although proposals have been made for biomarker study registration,¹⁰ resources to support registration systems are needed along with incentives or requirements from journals and funders, similar to existing mandates for registration of clinical trials in ClinicalTrials.gov.

CONCLUSION

A concerted effort involving many stakeholders is needed to provide guidance, resources, and incentives to successfully achieve research reproducibility goals. Signs of increased commitment to reproducibility are encouraging, but additional stakeholders will need to join the effort in order to succeed in changing the culture and improving reproducibility of early clinical biomarker research (see **Supplementary Table S2**).

Conflict of Interest. The author declared no conflicts of interest.

1. Landis, S.C. *et al.* A call for transparent reporting to optimize the predictive value of pre-clinical research. *Nature* **490**, 187–191 (2012).
2. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials; Board on Health Care Services; Board on Health Sciences Policy; Institute of Medicine: In Evolution of Translational Omics: Lessons Learned and the Path Forward. (Ed. Micheel, C.M., Nass, S., Omenn, G.S.) (National Academies Press, Washington, DC, 2012), 1–15. <http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx>.
3. McShane, L.M. *et al.* Criteria for the use of omics-based predictors in clinical trials. *Nature* **502**, 317–320 (2013).
4. McShane, L.M. *et al.* Criteria for the use of omics-based predictors in clinical trials: Explanation & elaboration. *BMC Med.* **11**, 220 (2013).
5. Leek, J.T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
6. Fleming, T.R. Clinical trials: Discerning hype from substance. *Ann. Intern. Med.* **153**, 400–406 (2010).
7. Pepe, M.S., Fan, J., Seymour, C.W., Li, C., Huang, Y. & Feng, Z. Biases introduced by choosing controls to match risk factors of cases in biomarker research. *Clin. Chem.* **58**, 1242–1251 (2012).
8. Gray, R.J. Weighted analyses for cohort sampling designs. *Lifetime Data Anal.* **15**, 24–40 (2009).
9. Dwan, K. *et al.* Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* **3**, e3081 (2008).
10. Andre, F. *et al.* Reporting cancer biomarker studies in their appropriate landscape: a call for a comprehensive biomarker study registry. *Nat. Rev. Clin. Oncol.* **8**, 171–176 (2011).

Published 2017. This article is a U.S. Government work and is in the public domain in the USA. Clinical and Translational Science published by Wiley Periodicals, Inc. on behalf of American Society for Clinical Pharmacology and Therapeutics.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Supplementary information accompanies this paper on the *Clinical and Translational Science* website. ([http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1752-8062](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1752-8062))