Statistical Principles for Omics-based Clinical Trials

Michael C Sachs

National Cancer Institute

maito:michael.sachs@nih.gov

October 2014. Incomplete Draft. Please do not cite without permission.

Abstract

High-throughput techologies enable the measurement of a large number of molecular characteristics from a small tissue sample. High-dimensional molecular information (referred to as omics data) offers the possibility of predicting the future outcome of a patient (prognosis) and predicting the likely response to a specific treatment (prediction). Embedded in the vast amount of data is the hope that there exists some signal that will enable practicioners to deliver therapy personalized to the molecular profile of a tumor, thereby improving health outcomes. The challenges are to determine that the omics assays are valid and reproducible in a clinical setting, to develop a valid and optimal omics-based test that algorithmically determines the optimal treatment regime, to evaluate that test in a powerful and unbiased manner, and finally to demonstrate clinical utility: that the test under study improves clinical outcome as compared to not using the test. We review the statistical con-

siderations involved in each of these stages, specifically dealing with the challenges of high-dimensional, omics data.

Keywords. genomics; personalized medicine; predictive biomarker; statistics

1 Introduction

Omics technolgies that generate a large amount of molecular data about a cancerous tumor have the potential to provide accurate predictions of a patient's prognosis and predictions of their response to a specific treatment regime. The idea of omics-based biomarkers is that distinct tumor types can be identified using the multi-dimensional molecular data leading to treatment decisions personalized to that tumor type. An omics-based test can guide the decisions to treat or not to treat and help identify the particular therapy most likely to work. The challenge is to identify and demonstrate definitively that the use of an omics-based test improves clinical outcomes in a patient population.

An omics-based test can be used to predict a patient's prognosis, which is their expected clinical outcome. A test that provides accurate predictions of prognosis, regardless of treatment, is referred to as a prognostic biomarker. A predictive omics test is one that accurately predicts disease outcomes with the application of specific interventions. Predictive markers are therefore useful for the selection among two or more treatment options. Statistically, a prognostic test is strongly associated with clinical outcome and a predictive test modifies the association between treatment and clinical outcome (interaction). High dimensional omics data can be used to identify specific

molecular targets as potential mechnisms for drug development, however the use of omics technologies for drug development is beyond the scope of this review.

The path from development to definitively evaluating an omics-based test for prognosis or prediction of treatment response is long and arduous. Often, the end goal is to develop a test suitable for use in a clinical trial for guiding treatment. The oncology literature is full of reports that develop and/or evaluate omics-based tools for prognosis and prediction. Developing a simple test based on high-dimensional omics data can be complex and often uses novel statistical methods. Definitive evaluation of a prognostic or predictive test is costly and rife with methodological pitfalls. We aim to review such issues, giving you the resources to ask the right questions when critically weighing the evidence presented in a report of an omics-based study. Ultimately, as a practicing oncologist the question is: "Is this omics-based test something I want to use to improve patient care?".

The long road to implementing a test in a practice starts with analytical validation, that is, demonstrating that the omics-based assay accurately and reproducibly measures the molecular quantities. After the assay performance is established comes the test development and preliminary evaluation. This involves reducing the high-dimensional data into a one-dimensional quantity that will be used to make a decision. This one-dimensional quantity is often a risk score: an estimate of the probability of a speicific clinical outcome. It is necessary to establish the clinical validity of this risk score, that is, demonstrate that the risk score is independently associated with clinical outcome. Care must be taken to completely separate the development of the risk score from the

evaluation, otherwise estimates can be optimistically biased. Finally, the risk score must be translated into a binary decision, often using a threshold. It remains to demonstrate that the use of the test to make this decision improves patient outcomes.

The following sections specify questions you should ask while reading a report of an omics-based clinical study. We review the importance of such questions, and common pitfalls to watch for. If you are reporting on an omics-based trial, answers to these questions should be made clear to the reader. Formal efforts to guide reporting have been developed, such as the REMARK checklist (1), the GRIPS statement (2), and a third guideline article that lacks an acronym (3). Our review reflects these efforts through the readers' lens.

2 Terminology

An omics-based test, or simple an **omics test**, is a mapping from the set of features on the omics assay to a single number. This number can be a binary value, such as good or poor prognosis, or it can provide a continuous scale, such as a risk score. It must be feasible to perform the test on an individual patient basis, by measuring the omics assay on the individual's tissue. The assay generates lots of measurements, which we will refer to as **features**, and then fixed mathematical calculations are done to tranform the many features into the single test value. Examples of such features are gene expression values, protein expression measurements, or genomic mutations.

Investigators determine the way in which the mathematical calculations in the **development phase**. Often, there is a complete sample which is randomly allocated

into roughly equal **development** and **validation** samples. These are also sometimes referred to as **training** and **test** sets of samples. A report may cover only one of the two steps. At the end of the development phase, the model for the mathematical calculations is fixed and locked down.

That model is evaluated definitively in the **validation** phase in a completely independent sample. In order for the validation to be unbiased and definitive, it is imperative that no information from the validation sample leaks into the development phase. The validation should mimic realistic clinical use as much as possible, and that means that no further refinement to the test is allowed based on the observed results.

3 What is the intended clinical use?

Do: define the clinical use (4)

As with all clinical studies, the end goal is to improve patient care. Omics studies are no different, and a clear statement of the intended clinical use of the omics-test should be prominent. Carefully describing the context for the use of the assay determines the type of study needed to develop and validate it. The intended use of the assay also provides an overarching context in which to interpret the population under study, the assay measurements, and the statistical methods.

4 What is the patient population of interest?

Along with the intended clinical use, a report should have a clear statement of the intended population in which the test is being evaluated. This could be broad or quite specific.

5 Is the omics assay valid?

Analytical validation of an assay involves evaluating the performance of the measurement in terms of accuracy, bias, and precision under a variety of conditions. Conditions are things like preanalytic factors such as specimen quality, specimen collection, storage, and processing procedures, and technical aspects such as laboratory technician and batch effects from reagent lots or other assay materials. The high-dimensional nature of omics data makes it very difficult to assess each of the hundreds or thousands of outputs from a single assay. In developing a omics-based signature that only uses a subset of the components of a high-dimensional assay, one can analytically validate the final signature alone. However, prior to developing the signature, one must develop detailed standard operating procedures for specimen handling and processing to ensure a baseline level of validity.

Did the authors of the report state what type of specimens were used in the study? Can the test be applied to formalin-fixed parafin embedded (FFPE) tissue, or only fresh-frozen? Most omics-based assays require a minimum percentage of tumor to be sucessful. A report should clearly state what criteria were used to screen tissue samples

prior to running the assay. Generally this involves a criteria for the rejection of poor-quality specimens on the basis of percent tumor, percent necrosis, or some other marker of tissue quality.

Molecular assays can sucessfully be run on decades old FFPE tissue (5). However, factors involved in the tissue processing and storage can impact the results (6–8). Due to the high dimensionality of omics assays, a small amount of bias on each feature can translate into large errors when incorporating data from hundreds or thousands of features into a single continuous measurement. Therefore it is important to assess the impact of processing on the individual features in addition to the overall test.

In addition to processing and storage, technical aspects of an assay can impact the final results in a predictable way (9,10). There could be technicial effects, differences due to reagent lots, and other batch effects. Such batch effects are commonly recognized yet often ignored in high-dimensional assays (11). Efforts should be made to measure the impact of these technical aspects and minimize them to the greatest extent possible. The way in which samples are assayed should be randomized to prevent confounding batch effects with the clinical outcome. Development and validation samples are sometimes run in the same batch or with the same lot of technical aspects. This does minimize batch effects, however, it can provide an overly optimistic assessment of the test, because in clinical use, running things in the same batch is not an option.

6 What does the omics-test do?

Does the test provide a continuous score or a binary classification?

How are the features of the omics assay translated into a clinically meaningful quantity?

Compare: feature filtering based on association with outcome, regularization. (12,13)

Do: consider all available methods, model averaging. Hard to determine best method in advance.

Don't: rely on clustering to yield good predictions of outcome.

7 On what samples was the test developed?

Similar to developing criteria for rejection of tissue samples, in omics settings, criteria should be developed for the rejection of individual features (e.g. genes, proteins) prior to the development of the test. Features that do not pass the pre-specified quality metrics should be removed from consideration from the final test. Note that this feature processing step does not involve any clinical outcome measurements. As a concrete example, in the development of a gene expression based test, investigators may choose to exclude probe locations that have a dynamic range under some threhold, or probles for which only a small proportion of the samples had calls, or probes that have absolute expression levels below some theshold. Quality control steps like this can ensure a more robust a reproducible development of the test.

Study design: consider retrospective (14)

Don't: confound technical factors with clinical outcomes. (11,15)

Do: maintain strict separation between development and evaluation.

Do: cross validation if you have a data-sparse setting. (16–19)

Don't: use convoluted methods leading to overfitting.

8 On what samples is the test being evaluated?

Do: define the clinical use (4)

Do: Design your study appropriately to answer the clinical question definitively (20–31)

Do: Power your trial appropriately (32,33)

Don't: do partial resubstitution

9 Are valid methods being used to evalute the test?

Bad: IDI or net reclassification (34,35)

Bad: Comparing AUCs for regression models (36)

Good: comprehensive and pre-specified approach (37)

10 Are the development and evaluation samples strictly separated?

This issue has come up in previous sections, yet this error occurs so frequently that it needs to be highlighted in its own section. The evaluation sample for the assessment of a prognostic or predictive test needs to be completely independent from the development sample. This is especially true for omics-based tests, whose developement is often complex and convoluted. Any imformation from the evaluation sample that leaks into

the development sample can bias the results, making tests appear better than they truly are.

Don't: do partial resubstitution

Don't: make these mistakes (38–40)

11 Concluding remarks

Do: follow reporting criteria (1–3,41)

12 References

1. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (rEMARK): Explanation and elaboration. BMC medicine. BioMed Central Ltd; 2012;10(1):51.

2. Janssens AC, Ioannidis J, Bedrosian S, Boffetta P, Dolan SM, Dowling N, et al. Strengthening the reporting of genetic risk prediction studies (gRIPS): Explanation and elaboration. European journal of clinical investigation. Wiley Online Library; 2011;41(9):1010–35.

3. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, et al. Criteria for the use of omics-based predictors in clinical trials: Explanation and elaboration. BMC medicine. BioMed Central Ltd; 2013;11(1):220.

4. Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: Lessons from real trials. Clinical trials. SAGE Publications; 2010;7(5):567–73.

5. Iwamoto KS, Mizuno T, Ito T, Akiyama M, Takeichi N, Mabuchi K, et al. Feasibility of using decades-old archival tissues in molecular oncology/epidemiology. The American journal of pathology. American Society for Investigative Pathology; 1996;149(2):399.

6. Srinivasan M, Sedmak D, Jewell S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. The American journal of pathology. Elsevier; 2002;161(6):1961–71.

7. Maldegem F van, Wit M de, Morsink F, Musler A, Weegenaar J, Noesel CJ van. Effects of processing delay, formalin fixation, and immunohistochemistry on rNA recovery from formalin-fixed paraffin-embedded tissue sections. Diagnostic Molecular Pathology. LWW; 2008;17(1):51–8.

8. Specht K, Richter T, Mller U, Walch A, Werner M, Hfler H. Quantitative gene expression analysis in microdissected archival formalin-fixed and paraffin-embedded tumor tissue. The American journal of pathology. Elsevier; 2001;158(2):419–29.

9. Pennello GA. Analytical and clinical evaluation of biomarkers assays: When are biomarkers ready for prime time? Clinical Trials. SAGE Publications; 2013;1740774513497541.

10. Isler JA, Vesterqvist OE, Burczynski ME. Analytical validation of genotyping assays in the biomarker laboratory. Future Medicine Ltd; 2007;

11. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. Nature Publishing Group; 2010;11(10):733–9.

12. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS biology. Public Library of Science; 2004;2(4):e108.

13. Hastie T, Tibshirani R, Friedman J, Hastie T, Friedman J, Tibshirani R. The elements of statistical learning. Springer; 2009.

14. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. Journal of the National Cancer Institute. Oxford University Press; 2009;101(21):1446–52.

15. Soneson C, Gerster S, Delorenzi M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. PloS one. Public Library of Science; 2014;9(6):e100335.

16. McShane LM, Polley M-YC. Development of omics-based clinical tests for prognosis and therapy selection: The challenge of achieving statistical robustness and clinical utility. Clinical Trials. SAGE Publications; 2013;10(5):653–65.

17. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio) marker. Heart. BMJ Publishing Group Ltd; British Cardiovascular Society; 2012;98(9):683–90.

18. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. external validation, model updating, and impact assessment. Heart. BMJ Publishing Group Ltd; British Cardiovascular Society; 2012;heartjnl–l2011.

19. Polley M-YC, Freidlin B, Korn EL, Conley BA, Abrams JS, McShane LM. Statistical and practical considerations for clinical evaluation of predictive biomarkers. Journal of the National Cancer Institute. Oxford University Press; 2013;105(22):1677–83.

20. Freidlin B, Korn EL. Biomarker enrichment strategies: Matching trial design to

biomarker credentials. Nature Reviews Clinical Oncology. Nature Publishing Group; 2014;11(2):81–90.

21. Baker SG, Sargent DJ. Designing a randomized clinical trial to evaluate personalized medicine: A new approach based on risk prediction. Journal of the National Cancer Institute. Oxford University Press; 2010;

22. Baker SG, Kramer BS, Sargent DJ, Bonetti M. Biomarkers, subgroup evaluation, and clinical trial design. Discovery medicine. 2012;13(70):187–92.

23. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, et al. Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. Statistics in medicine. Wiley Online Library; 2009;28(10):1445–63.

24. Denne JS, Pennello G, Zhao L, Chang S-C, Althouse S. Identifying a subpopulation for a tailored therapy: Bridging clinical efficacy from a laboratory-developed assay to a validated in vitro diagnostic test kit. Statistics in Biopharmaceutical Research. Taylor & Francis; 2014;6(1):78–88.

25. Eng KH. Randomized reverse marker strategy design for prospective biomarker validation. Statistics in medicine. Wiley Online Library; 2014;

26. Freidlin B, Jiang W, Simon R. The cross-validated adaptive signature design. Clinical Cancer Research. AACR; 2010;16(2):691–8.

27. Freidlin B, McShane LM, Polley M-YC, Korn EL. Randomized phase iI trial designs with biomarkers. Journal of Clinical Oncology. American Society of Clinical Oncology; 2012;30(26):3304–9.

28. Freidlin B, Korn EL, Gray R. Marker sequential test (maST) design. Clinical Trials.

SAGE Publications; 2013;1740774513503739.

29. Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. Journal of the National Cancer Institute. Oxford University Press; 2007;99(13):1036–43.

30. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. Journal of Clinical Oncology. American Society of Clinical Oncology; 2009;27(24):4027–34.

31. Morita S, Yamamoto H, Sugitani Y. Biomarker-based bayesian randomized phase iI clinical trial design to identify a sensitive patient subpopulation. Statistics in medicine. John Wiley & Sons, Ltd; 2014;

32. Mackey HM, Bengtsson T. Sample size and threshold estimation for clinical trials with predictive biomarkers. Contemporary clinical trials. Elsevier; 2013;36(2):664–72.

33. Peterson B, George SL. Sample size requirements and length of study for testing interaction in a 1×< i> k</i> factorial design when time-to-failure is the outcome. Controlled clinical trials. Elsevier; 1993;14(6):511–22.

34. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: Do not rely on integrated discrimination improvement and net reclassification index. Statistics in medicine. John Wiley & Sons, Ltd; 2013;

35. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. American journal of epidemiology. Oxford University Press; 2011;kwr013.

36. Seshan VE, Gnen M, Begg CB. Comparing rOC curves derived from regression models. Statistics in medicine. Wiley Online Library; 2013;32(9):1483–93.

37. Janes H, Brown MD, Huang Y, Pepe MS. An approach to evaluating and comparing biomarkers for patient treatment selection. The international journal of biostatistics. 2014;10(1):99–121.

38. Lee S. Mistakes in validating the accuracy of a prediction classifier in high-dimensional butsmall-sample microarray data. Statistical methods in medical research. SAGE Publications; 2008;

39. Sargent DJ, Mandrekar SJ. Statistical issues in the validation of prognostic, predictive, and surrogate biomarkers. Clinical Trials. SAGE Publications; 2013;10(5):647–52.

40. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of dNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute. Oxford University Press; 2003;95(1):14–8.

41. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: A systematic review. PLoS medicine. Public Library of Science; 2012;9(5):e1001221.