Study Design Considerations for Cancer Biomarker Discoveries

Yingye Zheng
Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N., M2-B500, Seattle, Washington 98109, yzheng@fhcrc.org, Phone: 206-667-7580, Fax: 206-667-7004

Abstract

Background—Biomarker discovery studies have generated an array of omic data, however few novel biomarkers have reached clinical use. Guidelines for rigorous study designs are needed.

Content—Biases frequently occur in sample selection, outcome ascertainment, or unblinded sample handling and assaying process. The principles of a prospective-specimen collection and retrospective-blinded-evaluation (PRoBE) design can be adapted to mitigate various sources of biases in discovery. We recommend establishing quality biospecimen repositories using matched two-phase designs to minimize biases and maximize efficiency. We also highlight the importance of taking the clinical context into consideration in both sample selection and power calculation for discovery studies.

Summary—Biomarker discovery research should follow rigorous design principles in sample selection to avoid biases. Consideration of clinical application and the corresponding biomarker performance characteristics in study designs will lead to a more fruitful discovery study.

Impact—Appropriate study designs will improve the quality and clinical rigor of biomarker discovery studies.

Keywords
bias; discovery study; two-phase designs; case-cohort study; nested case-control study; sample size; specimen repository

1. Introduction

Novel markers for outcome prediction hold the promise to dramatically change the decision-making process of modern medicine. For example, the blood level of prostate specific antigen (PSA) is used for prostate cancer screening as well as monitoring disease progression. However, due to poor specificity for detecting aggressive cancers, relying on PSA alone as a strategy for disease screening has major limitations and often leads to over-diagnosis and overtreatment of indolent cancers (1). Hence, there is a need for novel biomarkers to improve upon the limitations of current clinical biomarkers such as PSA. With
the recent technical advancement in laboratory science and dissemination of sophisticated bioinformatics tools, there is a great opportunity for the discovery of novel markers that ultimately lead to improvement in medical practice (2).

Biomarker development is a complex process, constituting several phases to produce a meaningful clinical tool (3). Specific development might go through phases differently but they can be in general regarded as a discovery or a validation study. The discovery study typically consists of a pre-clinical exploration with a pool of candidate biomarkers. The objective is to identify a short list of promising markers that are associated with clinical outcome of interests for further investigation. The reliability and validity of clinical assays are also established at this stage. For biomarkers that are intended for clinical use, an algorithm for combining a biomarker panel needs be further constructed and a decision rule including positivity threshold needs to be identified as part of the final phase in discovery, or ‘pre-validation’. The discovery research poses considerable challenges, due to a large number of biomarkers being investigated, the typical weak signal from an individual marker, and frequently strong noise due to experimental effects (4). The validation study is a key step for translating the laboratory findings into clinical practice. It includes both retrospective and prospective evaluation of a ‘locked down’ biomarker-assisted decision rule, in terms of its clinical utilities in the relevant population and the impact on the disease burden within the intended population. For each type of biomarker studies, special design considerations and integrated statistical approaches are needed for an efficient and successful pursuit.

While evolving Molecular technologies in discovery studies have generated an array of omic data, to date success has been very limited considering numbers of biomarker discoveries that have reached clinical use (5). As a number of researchers have recognized (5, 6, 7), a contributing reason to such a phenomenon is the lack of adequate study designs in discovery research. In contrast to biomarker validation studies, less attention has been demanded on the clinical rigor, because the studies often considered exploratory in nature. The use of hospital-based design or convenient samples is abundant in the literature. Such samples are often inherently biased, results in an increase in false positives and missed true signals. The bias often cannot be corrected even with a sophisticated statistical procedure. Indeed the highest quality biospecimen and carefully implemented experimental protocols will be wasted if the right samples are not used.

In this paper, we provide a review of study design strategies that may elevate the quality and efficiency of discovery research. We first describe sources of biases that frequently occur in biomarker research. We then discuss strategies for choosing biological samples to avoid systematic biases, drawn on the guidelines that are currently widely applied in validation research. We also discuss sampling considerations that can enhance the power of discovery studies.
2. Sources of Bias in Biomarker Discovery

Bias in biomarker research can occur at each step of biomarker development, from sample acquisition and handling, study participant selection, assaying process, to statistical analysis and interpretation of the results.

Many discoveries start with a selection of cases (with diseases presence or harboring a bad outcome) and controls (without the disease or harboring good outcome) to participate in the study. The case-control study allows the comparison of the prevalence of biomarkers between individuals with the outcome of interests versus these without the outcome. However, it is also prone to selection bias, which is the most common in biomarker study particular in the early discovery phase. Convenient samples are often used at this stage, due to both the exploratory nature of the study, and limited access to high quality biospecimen repositories that are strategically preserved for subsequent confirmatory or validation studies. In a case-control study, often there are confounding factors which are not specifically associated with disease status but only associated with certain characteristics of the diseased patients. For example, cases and controls may differ in characteristics such as age, gender and other biological parameters associated with the biomarkers under study (5, 8). In multi-center studies, samples from comparison groups may be collected at different clinics where there is discrepancy in population and sampling handling procedures. As a number of biomarkers are considered, initial screening often including a marginal comparison of the two groups (such as a two-sample t-test) without statistically controlling for confounding factors. Thus false discovery can be very severe at this stage due to selection bias.

Bias related to the ascertainment of study endpoint is another common problem in biomarker research. Cases and controls may not be directly corresponding to the clinical settings the biomarkers are intended for. Severe cases tend to be selected. For example, in pancreatic cancer, biomarkers are sought to detect the disease at early stage when curative treatment is possible. Using samples obtained from late-stage cases may lead to an overestimation of the performance of early detection biomarkers, and set up for failure in the downstream validation. Inappropriate selection of control subjects is also a major cause of systemic bias. In a prospective study, an individual with short follow-up period with no event yet being observed may be selected as a control. Without a further follow-up and assessment of the disease status, study may underestimate biomarkers’ capacity in discriminating between individuals with and without the disease. In other settings ascertainment bias may occur when a participant’s disease status is ascertained selectively based on known in-formation correlated with biomarker levels. For example due to the widespread of screening with PSA, an ascertainment of cancer status with biopsy is often performed on patients with an elevated PSA or changes in its velocity (9). Novel biomarkers related to PSA may subject to false discovery using samples obtained from a cohort with such an ascertainment scheme.

Bias can also occur during biomarker testing and statistical analysis. Cases and controls may differ in laboratory procedures such as storage duration and other unknown confounders (10, 11). Knowledge of the subject’s outcome may impact how the specimen is processed, and more importantly the interpretation of the assay result. When the goal of a discovery study is
to identify a biomarker panel, overfitting bias can be a major concern when individuals are included in both discovery set and validation set.

3. Design considerations that can mitigate potential bias

There are two common design options in population studies: the prospective and retrospective studies. In prospective study participants are followed for a clinical outcome, with exposure formation and biologic samples being collected at baseline or repeatedly over time. It has the advantage of minimizing potential biases and better capture of environmental exposures before disease onset and clearer characterization of traits that evolving over time (12). However due to the cost of assaying biomarkers, seldom a full prospective cohort can be employed for biomarker evaluation, in particular, if the outcome of interest is rare such as cancer. A retrospective study identifies comparison groups at the start of the study and looks backwards and examines risk factors in relation to the outcome. Most sources of error due to confounding and selection are common in retrospective studies. Therefore much deliberation is needed for biomarker study design, in order to conduct cost-effective research while maintaining the rigor of the study.

To provide standards of practice for conducting biomarker validation studies, Pepe et al. (13) proposed a prospective-specimen collection, retrospective-blinded-evaluation (PRoBE) design. Since the same sources of error are encountered in discovery studies, it is contended that the key principles involved in PRoBE design can be applied to the discovery studies to migrate the potential biases (7). A first key component of the design principle is the consideration of clinical context early in the study. Up-front a study should address the questions such as: what clinical settings will the biomarkers be applied to? Or, what patient population would benefit from the markers? Many clinical applications can be envisioned, for example, diagnosing patients who harbor disease at the time of testing (diagnostic markers), stratifying patients at different risks for a future outcome such as progression or death (prognostic markers), or selecting the most appropriate treatment regimens (predictive markers). Sampling then should be done on a cohort from the identified targeted population and study endpoints and data types (e.g., binary outcome versus time-to-event outcome) should then be considered accordingly. By using samples from the targeted population, discovery study will avoid extrapolating bias and would more likely produce results that can be successfully validated subsequently, and will improve on the “generalizability” of any marker subsequently found. The second principle of PRoBE design is that specimens and clinical data should be collected prospectively and prior to clinical outcome ascertainment. The prospective collection captures the advantages of a prospective study and ensures a uniform specimen collection for all subjects, eliminating systematic biases due to different selection and collection procedures between the comparison groups. A third design principle is that cases and controls should be retrospectively selected randomly from the cohort. This is a further step for eliminating potentially unmeasured confounding factors. Finally, PRoBE requires specimens assayed for biomarkers in a fashion that is blinded to clinical outcome measures to migrate differential interpretations of testing results.

The PRoBE design involves constructing a case-control study nested within a prospective cohort, a type of two-phase sampling designs originally proposed for assessing exposure-
disease association in an assembled cohort. The designs have also been recognized as resource-efficient sampling strategies for molecular marker studies (14, 15). Two-phase sampling designs including the nested case-control (NCC) and case-cohort (CCH) studies (16, 17). In an NCC study, all individuals observed to have a clinical event are selected as cases. At each selected case’s failure time, a random sample of a few individuals is selected without replacement among those who are still being followed (the risk set of the cases) as potential controls. Biomarkers then measured on the selected cases and controls. In a case-cohort study, biomarker study is conducted on all participants with events and a subcohort that is randomly sampled from the entire cohort. The advantages of the two-phase designs include preserving the prospective collection of biological samples years before clinical diagnosis or progression, a clear definition of the source population from which to select controls. They also accommodate issues with lost to follow-up in a prospective cohort.

Matching is a very useful strategy for eliminating confounding factors. In discovery studies, sample sizes are often limited to allow for stratification and adjustment for multiple confounders in a regression model. However, matching can be applied in study design as well as in data analysis. For example, matching biological samples for differences in handling, processing and duration of storage can be easily handled in both CCH and NCC designs. The NCC design essentially matches cases and controls on the length of follow-up, therefore, matches the comparison groups in terms of storage times and other unmeasured time-related confounding factors (15). The analysis of matched data in the context of association parameters is well developed, however, interpretation and analysis of biomarker performance with matched data need to be carefully considered (18, 13). The analysis of matched data under two-phase sampling has been recently developed in the context of biomarker evaluation (19, 20, 21). Strategies for sampling specific subsets in order to improve study efficiency has also been considered specific to biomarker studies. For example, (15) suggested a ‘counter-matching’ approach and (22) evaluated the impact of different two-phase matched sampling strategies on the efficiency for calculating biomarker performance.

4. Design considerations that can enhance the power of a discovery study

Discovery studies often aim to explore a haystack of candidate markers and select for more careful investigation a subset that is potentially useful in clinical practice. Sometimes a more ambitious goal is to further develop a combination algorithm of a panel of biomarkers from a large number of candidate biomarkers. Such tasks pose considerable statistical challenges, yet often are addressed with a small number of samples.

Sufficient sample sizes are needed to make reliable recommendations about which set of markers should be filtered for further analytical and clinical validation. There has been much research into formal multiple hypothesis testing procedures for designating biomarkers as differentially expressed (23). Procedures that control error rates have been the main focus of statistical research. The false discovery rate (FDR) is the expected proportion of “discoveries” (rejected null hypotheses) that are false (incorrect rejections). Sample sizes can be determined based on a pre-specified criterion that controls such an error. The other popular approach is based on the ranking of the statistics, such as t-statistics or odds ratio,
and the selection of the top \( k \) performers, where \( k \) is often decided based on the cost constraint for further evaluation and assay development. When pilot studies are available, sample sizes can be determined by simulation procedures that take into account the multiplicity and dependences encountered in discovery data (24).

One limitation of the standard procedures for filtering and ranking biomarkers is that they do not quality downstream performance in a clinically relevant way. This gap between discovery and validation may partly contribute the failure of validating ‘promising’ biomarkers. To improve the clinical rigor of discovery research, a critical preliminary step is to specify the clinical performance criteria a successful marker will ultimately achieve. Then the discovery can be conducted by searching directly markers that will enhance the intended clinical outcome and use enough of samples to ensure the informative markers will be selected. Tying the design of discovery study to the clinical application, (7) proposed an alternative approach for sample size determination drawn on the notions of validation but adapted to the context of biomarker discovery research. The approach requires first the specification of a measure for biomarker performance, and defining informative and useless markers accordingly. Sample sizes can then be gauged by two quantities: discovery power, the proportion of useful markers the study should identify, and false leads expected, the tolerable number of useless markers among these identified. The approach is useful in the identification of clinically relevant biomarkers.

5. Illustration

We illustrate the approach with a prostate cancer urine biomarker studies conducted at the National Cancer Institute’s Early Detection Research Network (EDRN). The study objective is to identify novel biomarkers to develop a risk prediction tool combing biomarkers with other clinical measures to predict Gleason score upgrade from 6 on biopsy to Gleason 7 and above on subsequent radical prostatectomy (RP). The intended clinical application is to use clinical predictors and biomarkers (e.g., PCA3, TMPRSS2:ERG and other selected markers) to accurately predict the existence of Gleason 3+4 or higher disease confirmed at radical prostatectomy (25). The primary endpoint of this study is the presence of tumor upgrading at the time of radical prostatectomy.

The study will recruit a prospective cohort of men presenting with biopsy Gleason 3+3 prostate cancer who have elected to undergo radical prostatectomy for prostate cancer. Preoperative blood and urine samples, prostate biopsy tissue samples, as well as clinical and demographic data, will be collected. A protocol is developed for enrollment, specimen processing and storage and will be implemented across study centers for uniform collection. Cases and controls will be determined after the RP, and their status will be confirmed by a central pathology review. The outcome is not known at the time of sample collection and will be blinded to laboratories by labelling specimens with randomly generated identification numbers. Samples will be ordered randomly for processing. The study, therefore, follows the PRoBE principles. Individuals will contribute samples to two independent sets, one reserved for biomarker discovery, the other will be used for clinical validation.
To determine the sample size for the discovery set, we consider the performance criteria as specificity evaluated at sensitivity fixed at 98%. The requirement of high sensitivity at 98% is due to the malignant nature of the outcome and a missing rate of 2% is considered tolerable. We then define a strong marker as having a specificity of 35% given 98% sensitivity, or equivalently, a negative predictive value (NPV) equals 0.92 with a 60% upgrading rate. This is the same criterion held for subsequent validation. We also define an intermediate marker as having a specificity of 15% at 98% sensitivity (NPV = 0.83), and an uninformative marker as having specificity = 0.02% at 98% sensitivity (i.e., NPV = 0.40).

For discovery study, identifying intermediate markers are useful, as they can be used in a panel to achieve the higher performance required at the validation stage. The null hypothesis is that specificity at 0.98 sensitivity is less than 2%. Since the goal is to identify useful markers for a panel, we consider the detection power for sample size calculation using the procedure described in details in (7). Using 10000 simulated samples drawn from populations with binormal ROC curve defined by the above performance criteria for a strong, an intermediate and an uninformative marker respectively, with a sample size of 200, 100% of such strong markers will be detected, 93% of intermediate markers will be identified.

6. Summary

Rigorous and efficient study design is critical to a fruitful discovery study. In particular, exploration should be based on study samples judiciously selected to avoid biases. Widely adopted strategies in validation studies, including prospective-specimen collection and retrospective-blinded-evaluation, should be considered in discovery studies to mitigate various sources of biases. Good quality biospecimen repositories should be made available for discovery research. The rigor of biomarker discovery study can be improved by putting effort and resources to set up reference sets following key design principles, similarly to what has been carried out in the EDRN for validation study (26). It is recognized that laboratory scientists are often well tuned in dealing with pre-analytical variations under controlled experimental conditions, however, are yet to gain an appreciation for sources of bias derived from conducting population studies in observational settings (6, 7). The field will benefit from engaging population scientists and biostatisticians in collaboration in the early exploration stage.

Abbreviations:

| Abbreviation | Description |
|--------------|-------------|
| PRoBE        | prospective-specimen collection and retrospective-blinded-evaluation |
| PSA          | prostate specific antigen |
| NCC          | nested case-control |
| CCH          | case-cohort |
| EDRN         | Early Detection Research Network |
| RP           | radical prostatectomy |
References

1. Barry MJ. Screening for prostate cancer — the controversy that refuses to die. N Engl J Med 2009;360:1351–4. [PubMed: 19297564]

2. Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nat Clin Pract Oncol 2008;5:588. [PubMed: 18695711]

3. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93:1054–61. [PubMed: 11459866]

4. Qin L-X, Levine DA. Study design and data analysis considerations for the discovery of prognostic molecular biomarkers: a case study of progression free survival in advanced serous ovarian cancer. BMC Med Genomics 2016;9:27. [PubMed: 27282150]

5. Diamandis EP. Cancer biomarkers: can we turn recent failures into success? J Natl Cancer Inst 2010;102:1462–7. [PubMed: 20705936]

6. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 2005;5:142. [PubMed: 15685197]

7. Pepe MS, Li CI, Feng Z. Improving the quality of biomarker discovery research: the right samples and enough of them. Cancer Epidemiol Biomarkers Prev 2015;24:944–50. [PubMed: 25837819]

8. McIntosh M, Anderson G, Drescher C, Hanash S, Urban N, Brown P, et al. Ovarian cancer early detection claims are biased. Clin Cancer Res 2008;14:7574.

9. Tangen CM, Goodman PJ, Till C, Schenk JM, Lucia MS, Thompson IM, Jr. Biases in recommendations for and acceptance of prostate biopsy significantly affect assessment of prostate cancer risk factors: Results from two large randomized clinical trials. J Clin Oncology 2016;34:4338–44.

10. Baker SG. Improving the biomarker pipeline to develop and evaluate cancer screening tests. Journal Natl Cancer Inst 2009;101:1116–9.

11. McLerran D, Grizzle WE, Feng Z, Bigbee WL, Banez LL, Cazares LH, et al. Analytical validation of serum proteomic profiling for diagnosis of prostate cancer: sources of sample bias. Clin Chem 2008;54:44–52. [PubMed: 17981926]

12. Manolio TA. Cohort studies and the genetics of complex disease. Nat Genetics 2009;41:5. [PubMed: 19112455]

13. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst 2008;100:1432–8. [PubMed: 18840817]

14. Rundle AG, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. Cancer Epidemiol Biomarkers Prev 2005;14:1899–1907. [PubMed: 16103435]

15. Rundle A, Ahsan H, Vineis P. Better cancer biomarker discovery through better study design. Eur J Clin Invest 2012;42:1350–9. [PubMed: 22998109]

16. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986;73:1–11.

17. Thomas DC. Addendum to “Methods of cohort analysis: Appraisal by application to asbestos mining”. J R Stat Soc Series A, General 1977;140:483–5.

18. Janes H, Pepe MS. Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. Biometrics 2008;64:1–9. [PubMed: 17501939]

19. Liu D, Cai T, Zheng Y. Evaluating the predictive value of biomarkers with stratified case-cohort design. Biometrics 2012;68:1219–27. [PubMed: 23173848]

20. Cai T, Zheng Y. Evaluating prognostic accuracy of biomarkers in nested case-control studies. Biostatistics 2011;13:89–100. [PubMed: 21856652]

21. Zheng Y, Cai T, Pepe MS. Adopting nested case-control quota sampling designs for the evaluation of risk markers. Lifetime Data Anal 2013;19:568–88. [PubMed: 23807695]

22. Zheng Y, Brown M, Lok A, Cai T. Improving efficiency in biomarker incremental value evaluation under two-phase study designs. Ann Appl Stat 2017;11:638–54. [PubMed: 28943991]
23. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci 2003;18:71–103.

24. Skates SJ, Gillette MA, LaBaer J, Carr SA, Anderson L, Liebler DC, et al. Statistical design for biospecimen cohort size in proteomics-based biomarker discovery and verification studies. J Proteome Res 2013;12:5383–94. [PubMed: 24063748]

25. Wei JT, Feng Z, Partin AW, Brown E, Thompson I, Sokoll L, et al. Can urinary pca3 supplement PSA in the early detection of prostate cancer? J Clin Oncology 2014;32:4066–72.

26. Feng Z, Kagan J, Pepe MS, Thornquist M, Rinaudo JA, Dahlgren J, et al. The early detection research network’s specimen reference sets: paving the way for rapid evaluation of potential biomarkers. Clin Chem 2013;59:68–74. [PubMed: 23193062]