Published in final edited form as:
Stat Med. 2012 September 28; 31(22): 2513–2515. doi:10.1002/sim.5499.

The biomarker revolution

Enrique F. Schisterman[a,*,†] and **Paul S. Albert**[b]

[a]Epidemiology Branch, Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Boulevard Room 7B03, Bethesda, MD 20892-7510, U.S.A.

[b]Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Boulevard Room 7B05, Bethesda, MD 20892-7510, U.S.A.

The use of biomarkers to assess exposures and investigate biomedical questions is exploding in epidemiological and medical research. Exploring the relationships between biomarker levels and health outcomes can have potentially profound effects on the biomedical community, leading to new etiological discoveries, as well as increased diagnostic capabilities for disease.

In estimating distributional parameters for a particular biomarker, population scientists face two main challenges: (1) overcoming the cost of performing a large number of assays and (2) dealing with measurement error either due to technical or intra-individual variability. In terms of cost, evaluating biomarkers in epidemiological studies can not only be expensive but time consuming as well. The power gained by a large sample must be weighed against the cost of performing more assays. Instrument sensitivity may also be problematic when studying levels of certain biomarkers inducing measurement error. Some members of the population may have serum levels below a detection threshold. Under these circumstances, values at or above the detection threshold that is designated are measured and reported, but values below the detection threshold are unobservable, limiting the information one can utilize in his or her analysis. Even after reproducibility and variability are established for a biomarker, financial constraints often limit further evaluation to small sets of samples.

Many investigators use techniques such as random sampling or pooling biospecimens in order to cut costs and save time on experiments. To pool, two or more specimens are physically combined into a single ‘pooled’ unit for analysis. Thus, a greater portion of the population is assayed for the same price, resulting in the potential for more information being obtained with fewer assays. Additionally, pooling the specimens reduces the effective variance of the biomarker. Commonly, analyses based on pooled data can only be performed under strong distributional assumptions on the original data (unpooled), which are challenging to validate because the biospecimens are pooled. However, even if the cost of a single assay is not a major restriction in evaluating biomarkers, pooling can be a powerful

[*]Correspondence to: Enrique F. Schisterman, Epidemiology Branch, Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Boulevard Room 7B03, Bethesda, MD 20892-7510, U.S.A.. [†]schistee@mail.nih.gov.

design that increases the efficiency of estimation based on data that is censored because of an instrument's lower limit of detection. On the other hand, random sampling provides data that can be easily analyzed and provides an efficient estimator of the variance. However, random sampling methods are not optimally cost-efficient designs for estimating means, and pooling designs are not optimal designs for estimating the variances.

Although approaches have been suggested to deal with cost and measurement error issues individually, hybrid pooled–unpooled designs offer a strategy that can address both problems simultaneously. These hybrid designs involve taking a sample of both pooled and individual specimens.

Although great progress has been made in the area of pooled biospecimens, one may need special statistical methods to analyze this type of data. For example, Weinberg proposed a set-based logistic regression to analyze a hybrid design [1]. Danaher proposed a novel pooling strategy, where pooling is carried out within disease and environmental status to increase statistical power to test for gene–environment interactions [2]. In the longitudinal setting, Malinovsky proposed a Gaussian random effects model for efficient maximum likelihood estimation of variance components, particularly the intraclass correlation coefficient [3]. For evaluation of the discriminating ability of biomarkers, Schisterman and colleagues developed ROC analyses based on pooled samples to analyze this type of data [4–6]. However, there are other important types of analyses that need to be developed before these types of study designs and data analyses become mainstream.

The heavy use of distributional assumptions is one of the main constraints that researchers have faced. The use of repeated measures is a gold standard approach to tackle measurement error issues of biomarkers. Vexler and colleagues demonstrated that the hybrid design, coupled with estimation using empirical likelihood, can overcome the need for replications and lessen the parametric distributional assumptions while maintaining efficiency [7]. Their proposed methods should be extended to other settings, such as other linear and nonlinear regression settings.

Whitcomb and colleagues extended the set-based logistic regression models proposed by Weinberg, relaxing the assumptions regarding exposure distribution and constant odds ratios that are often violated in practice [8]. They used the underutilized and flexible gamma distribution as a model assumption to overcome these constraints. Under the hybrid design, they showed that one can explicitly test these distributional assumptions. It remains to be explored what the optimal sample size of the pooled and unpooled biospecimens would be to efficiently evaluate the distributional assumptions.

Another type of biomarker commonly encountered in epidemiology and medicine is one where the result is the presence or absence of a trait, such as the presence or absence of a genetic abnormality, pregnancy, or HIV status. Such results are usually linked to a disease status. Lyles considered the case in which specimens are combined for the purpose of determining the presence or absence of a poolwise exposure, rather than assessing the actual binary exposure status for each member of the pool [9]. This is particularly clever because it maximizes the information of positive pools and augments with information from negative

pools. They provided a maximum likelihood approach for longitudinal studies where the exposures and outcomes are assessed multiple times. This approach remains to be extended to a combination of ordinal and continuous variables and has potential applications in many areas, including screening for genetic disorders and metabolomics, while accounting for possible confounders.

Pooled exposure measurements are generally assumed to be an average or, in some cases, a weighted average across subjects. However, this assumption might be violated if the unequal volume is pooled and then measured. This violation will not only affect the estimate of the mean but also the errors. Zhang and colleagues tackled this problem in the setting where the pools are dichotomized [10]. They provided an original method to estimate parameters in a logistic regression setting where a binary exposure is subject to pooling and the pooled measurement is dichotomized. They showed that those involving the exposure subject to pooling can be recovered from a poolwise binary regression model relating the pooled exposure measurements to the outcome and other covariates. They found that 'smart' pooling (i.e., pooling by important confounders), as has been shown before by Schisterman *et al.* in different settings [11], results in improved efficiency as compared with pooling by case status alone. This shows once again that understanding the causal relationships between variables, and therefore the potential confounders and effect modifiers, leads to more efficient and economical designs. The interplay between statisticians, basic biologists, and epidemiologists is paramount for understanding the complexities of the biomarker measurement, study design, and analysis in order to answer important etiological questions of interest.

The biomarker revolution is 20 years old and is not slowing down. Much has been done to move this field forward, but we have a long way to go. Every day, smart people out there are developing new and better ways to quantify pathophysiological processes and exposures. In our opinion, the use of hybrid designs for the evaluation of biomarkers is the most promising, cost-effective way of discovery for the future. Given the complexities of the measurement process, the underlying physiological pathways they measure, and the exponential rate of analytes that can be quantified simultaneously, we need to move forward from the traditional case point of individually measured biomarkers to creative new study designs and analytic techniques. There are always going to be expensive biomarkers, as new technology is always expensive. New discovery is hard, so if we want to be analyzing large numbers of expensive biomarkers, this is the only way. So let's get our hands dirty and join the biomarker revolution.

References

1. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. Biometrics. 1999; 55:718–726. [PubMed: 11314998]
2. Danaher MR, Schisterman EF, Roy A, Albert P. Estimation of gene-environment interaction by pooling biospecimens. Statistics in Medicine. 2012
3. Malinovsky Y, Albert PS, Schisterman EF. Pooling designs for outcomes under a Gaussian random effects model. Biometrics. 2012; 68:45–52. [PubMed: 21981372]
4. Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. Statistics in Medicine. 2003; 22:2515–2527. [PubMed: 12872306]

5. Mumford SL, Schisterman EF, Vexler A, Liu A. Pooling biospecimens and limits of detection: effects on ROC curve analysis. Biostatistics. 2006; 7:585–598. [PubMed: 16531470]
6. Vexler A, Schisterman EF, Liu A. Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. Statistics in Medicine. 2008; 27:280–296. [PubMed: 17721905]
7. Vexler A, Tsai W, Malinovsky Y. Estimation and testing based on data subject to measurement errors: from parametric to non-parametric likelihood methods. Statistics in Medicine. 2012
8. Whitcomb B, Perkins N, Zhang Z, Ye A, Lyles R. Assessment of skewed exposure in case-control studies with pooling. Statistics in Medicine. 2012
9. Lyles R, Tang L, Lin J, Zhang Z, Mukherjee B. Likelihood-based methods for regression analysis with binary exposure status assessed by pooling. Statistics in Medicine. 2012
10. Zhang Z, Lui A, Lyles R, Mukherjee B. Logistic regression analysis of biomarker data subject to pooling and dichotomization. Statistics in Medicine. 2012
11. Schisterman EF, Vexler A, Mumford SL, Perkins NJ. Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. Statistics in Medicine. 2010; 29:597–613. [PubMed: 20049693]