PhUSE US Connect 2018

Paper AB05

A Strategy for Evaluating Biomarkers Based on Emerging Technologies Using a Measurement Error Framework

Dong Wang, Ph.D.
Division of Bioinformatics Biostatistics,
National Center for Toxicological Research
US Food and Drug Administration

DISCLAIMER

The opinions expressed in this paper are those of the author and do not necessarily reflect the views of the US Food and Drug Administration.

ABSTRACT

With the rapid development of new biomarkers for precision medicine applications, there is a critical need for effectively incorporating new technologies into biomarker identification, treatment selection, and effect quantification. However, pushing at the boundaries of technological performance necessitates accepting a certain level of error rates for biomarkers based on technologies like deep sequencing. In this presentation, we outlined a new strategy in adapting the measurement error model framework to investigate the impact of performance characteristics of new technologies on major aspects of precision medicine applications and discussed future directions. This provides opportunities to map out the general boundaries where certain error rates can be tolerated for new biomarkers to still be effective. It can provide insight on how to conduct biomarker related studies at different stages while balancing efficiency and cost considerations.

This is a companion to my presentation at PhUSE US Connect 2018.

INTRODUCTION

We are in an era of rapid advancement of precision medicine, which has had wide ranging impacts on both patient care and regulatory considerations (Deng and Nakamura, 2017; Hyman et al., 2017). Probably the most prominent development is in cancer treatment. Since the breakthrough of the use of the kinase inhibitor imatinib for the treatment of chronic myelogenous leukemias (CML) that harbor the BCR-ABL1 balanced chromosomal translocation that greatly improved the outcome of this previously lethal form of leukemia (Druker et al., 2006), advancement has been made in multiple directions, including therapies targeting HER2, EGFR, and other cancer driver genes. Companion tests have become one of the hallmarks of the new generation of cancer or anti-viral drugs. All these developments require accurate and efficient identification and utilization of biomarkers, as well as classification of patients into meaningful subpopulations, which has been challenging with traditional technology (Vargas and Harris, 2016).

The development of new technologies in this area is equally breathtaking (Sheridian, 2017), with next generation sequencing (NGS) based approaches being the workhorse of precision medicine applications. The NGS technology makes it possible to obtain very high sequencing coverages for segments of the genome at a reasonable cost. It is now feasible to identify allele variants at low frequencies, which is often the case for somatic mutations. This, coupled with new procedures like liquid biopsy, has the potential to revolutionize clinical practices (Krishnamurthy et al., 2017). However, all sequencing platforms have their inherent error rates. For example, error rate as low as 0.1% has been reported for Illumina, which still might not be sufficient for some applications. Various new methods based on barcoding and single cell sequencing can achieve even better performance (Yang et al., 2017). But since it is desirable in a lot of cases to detect harmful mutations as early as possible, especially with highly diluted samples like those in liquid biopsy, we will always be pushing the edge of the performance limit of these technologies. Moreover, as human ages, the body accumulates random somatic mutations due to biological processes and environmental insults (Hoang et al., 2016). The variation between experimenters is another factor of concern (Torga & Pienta, 2017).

Some of these challenges have been discussed in the FDA discussion paper titled Optimizing FDA's Regulatory Oversight of Next Generation Sequencing Diagnostic Tests—Preliminary Discussion Paper (FDA, 2016). These include the lack of standard performance metrics for NGS based tests, the potential of detecting a very large number of variants, the difficulty to obtain valid clinical significance, and the difficulty to communicate the information to physicians and consumers. Fortunately, some of these issues are being actively addressed by efforts both inside and outside of FDA. One important project in this area is Project SEQC2 led by National Center for Toxicological Research with collaboration with other FDA centers and outside researchers (Shi et al., 2017). SEQC2 aims to develop standard analysis protocols and quality control metrics for consistent use of NGS data to enhance regulatory science research and precision medicine. Especially, it will provide a much needed evaluation for performance characteristics (specificity, sensitivity, reproducibility, optimal procedures) for the identification of somatic mutations as well as to assess its dependency on bioinformatics and coverage. It includes an extensive study with carefully designed

samples based on a number of cell lines that mimic settings for liquid biopsy, thus technical performance metrics can be readily derived. Related efforts also exist outside of FDA.

However, even with a better understanding of technical performance metrics, a large amount of work is still needed to clarify the implication to precision medicine applications (biomarker identification and validation, subpopulation identification, treatment selection, effect quantification). First, there is a diverse array of new technologies with varying performance under different settings. For example, for barcoding based NGS alone, there are bottleneck sequencing (BotSeqS), targeted error correction sequencing (TEC-Seq), Firefly NGS, and many other variations. Adding to this diversity is digital PCR based methods and combinations with protein based markers (Cohen et al., 2018). Also, these technologies have been applied in a diverse array of clinical settings using both traditional needle biopsy and liquid biopsy. Liquid biopsy itself includes tests focusing on circulating cancer cells and circulating cancer DNA (Krishnamurthy et al., 2017). Recently, tests using urine, saliva, and peritoneal fluid have also entered discussion. In reality, the accuracy in actual applications is often quite different from the optimal technical limit and varies significantly across different tissue types and ages (Bardelli and Pantel, 2017; Yang et al., 2017). Thus, results from standardized technical evaluations need to be put into proper context of applications. The understanding of these issues will be necessary to properly evaluate deep sequencing based applications in precision medicine.

In this presentation, we shall outline a strategy to approach the problems mentioned so far. We will discuss resources and methods that are already available and describe future directions for adapting existing statistical methods for the special needs in this area.

AVAILABLE RESOURCES

Fortunately, we can leverage significant work that has already been done in related areas. Researchers both inside and outside of FDA have some well recognized statistical approaches for each step of precision medicine development, including biomarker evaluation, subgroup identification, and clinical utility assessment. Excellent reviews can be found in Chen et al. (2014, 2015), Ondra et al. (2016), among others. In the traditional setting, one would assume that biomarker measurements (genomic markers, proteins, metabolites, etc.) for each patient can be obtained error free. This is often reasonable for biomarkers in clinical use previously. However, in using deep sequencing to detect somatic mutations, where we necessarily operate at the edge of the performance limit, taking into account of the error rates (and possibly biological background mutation rates) would be necessary to provide a precise evaluation for the utility of deep sequencing derived biomarkers or tests for applications in precision medicine. As mentioned in the FDA discussion paper (FDA, 2016), an important regulatory challenge in this area is to assure the safety and clinical validity of any test based on new technology while at the same time to allow consumers timely access to new tests. Thus while it is unreasonable to require all tests to be as accurate as in the traditional setting (essentially 100%), the measurement error issue has to be handled in

a disciplined and consistent fashion. There is a rich literature on measurement error problems in various scientific fields (Carroll et al., 2006 ; Yi, 2017). It is widely acknowledged that naively ignoring measurement errors can lead to severe bias and invalid conclusions. The impact of measurement errors can be quantified and the bias can be reduced in certain setting with a number of methods developed for a wide range of applications. An example of approaches dealing with measurement errors is the simulation extrapolation method (SIMEX; Cook & Stefanski,1994; Carroll et al., 2006), which has enjoyed wide applications due to its intuitiveness and easiness for implementation. Now, we shall discuss these issues in some detail.

Performance Metrics

First, we consider the technical performance metrics of deep sequencing based tests. These are performance characteristics on known and well characterized standard samples under well controlled experimental conditions. Since the importance of performance standard has been long recognized, publications for novel testing methods commonly include some characterization of performance metric with clearly defined experimental samples. Sensitivity, specificity, accuracy, and some other measures of performance are often reported. One task will be to catalog these technical performance measures for important testing methods. An overarching view of major technologies in this field will provide the basis for regulatory considerations even when new approaches come online. For this purpose, the SEQC2 project will be immensely valuable, which includes detailed evaluation of several important sequencing platforms (NextSeq, HiSeq, NovaSeq, Ion Torrent, MiSeq, etc.) with well characterized test samples. Besides synthetic samples with cell lines, it also includes two significant collections of trio (parents and child) samples, providing a valuable standard for internal verification. A set of neuroblastoma samples will also be used. SEQC2 is evaluating different sequencing applications (whole genome sequencing and targeted sequencing), coverage levels, and bioinformatics pipelines at multiple research sites. In addition to usual metrics like sensitivity/specificity, SEQC2 especially provides detailed evaluations of intra-site and inter-site variability. With SEQC2 results becoming available, it will provide important information for the review of technical performance metrics. Other comprehensive studies like that carried out by Association of Biomolecular Resource Facilities will also be useful, so are original research publications with performance metrics for important approaches. As these publications are from authors of differing backgrounds, the terminology and reporting methods could be quite different and sometimes confusing. So an important part of this exercise is to standardize the diverse results that are collected on the common statistical language.

As mentioned earlier, the technical performance is only one factor for deep sequencing based tests in precision medicine applications. In fact, it can be considered to define the upper limit of the performance. The third focus of the review will be on the application-wise performance metrics of various testing approaches. For realistic applications, one limitation is the amount of DNA, as both needle biopsy and liquid biopsy will only yield a limited amount of sample for analysis. This will put a limit on the achievable sequencing depth. Also especially for liquid biopsy, the circulating tumor DNA (ctDNA) is overwhelmed by free circulating DNA with normal cell origins. The stage of cancer as well as the tissue type also significantly affects the amount of ctDNA. Even for studies with high technical performance, the actual sensitivity of detecting

ctDNA in the plasma of cancer patient ranges widely from 50% to 95% across cancer stages and tissues. Researchers utilize a wide range of technologies in this area and sometimes combine sequencing based biomarkers with protein or other clinical biomarkers. The recent breakthrough of using other samples like urine and peritoneal fluid further complicates the picture. Another issue is regarding normal biological mutations that tissues invariably accumulate as people age (Krimmel et al., 2006). Though this does not yet pose a problem for the majority of applications so far, it has to be taken into consideration as the technology becomes even more sensitive. However, as the performance of deep sequencing based tests in a number of recently reported studies tend to reside in a common range (50%~95% sensitivity, 80%+ specificity), some general trends can already be observed by reviewing current literature.

Available Statistical Methods

There is a vast literature for statistical methods regarding biomarker related applications, which covers important steps of precision medicine practices: biomarker validation, subgroup identification, clinical utility assessment, among others. Applications in traditional settings usually cover a single or a small number of biomarkers, drawing on various types of regression and classification models as well as common inference frameworks. As the NGS and other high throughput technologies have make it possible to obtain a large number of potential biomarkers simultaneously, there is a flood of development for methods with a “machine learning” flavor that brings a unique set of challenges. In general most methods assume that the biomarker status can be obtained exactly. This is reasonable in the traditional biomarker setting. But as argued earlier, a certain level of error will be unavoidable for a lot of applications using deep sequencing and related technologies, which we want to address here.

Various approaches have been proposed in a number of scientific fields to deal with measurement errors in the data (Carroll et al., 2006; Yi, 2017). In the simple linear regression setting, the classical additive error and the Berkson models have been extensively studied, and the attenuation effect on naïve estimators of regression coefficients is well understood. However, for multiple linear regression and nonlinear models, the effect of measurement errors is more complex, possibly changing the magnitude or the signs of the estimate as well as resulting in incorrect coverage properties of the confidence interval. The measurement error effect for high dimensional models and machine learning type approaches have just begun to be explored in detail. These effects are often difficult to predict other than that it could lead to wrong conclusions when measurement errors are ignored.

For better inference under measurement error conditions, various approaches have been developed by many authors. Main approaches include regression calibration, simulation extrapolation (SIMEX), likelihood-based correction methods, unbiased estimating function methods, and Bayesian methods. Though excellent tutorials and reviews already exist for these approaches (Gustafson, 2004; Carroll et al., 2006; Buonaccorsi, 2010; Yi, 2017), it is still worthwhile to produce a specialized review with an eye on applications regarding deep sequencing based biomarkers. The reason is that the amount of literature is huge with methods developed for a wide range of scientific problems. Of these, only a subset would be relevant to the application to precision medicine. For example, for the problem we are interested in, an error free

validation sample usually would not be feasible, thus limiting the use of a subset of methods utilizing validation samples.

Adaptation of Statistical Methods

Here we outline a strategy for adapting existing statistical methods for dealing with measurement errors in the context of deep sequencing based biomarkers. We shall first focus on commonly used basic biomarker identification method (termed Univariate Regression by some authors). Methods dealing with a number of biomarkers simultaneously will be discussed later. Here "univariate" means either there is only one candidate biomarker (based on deep sequencing) to be used to select subpopulations at a time, or a composite variable has already been constructed from a biomarker panel and is the only variable to be considered to define subpopulations. There are several different designs for clinical trials to evaluate biomarkers in this setting, these have been discussed in several excellent reviews (Chen et al., 2014, 2015; Ondra et al., 2016). To illustrate, we consider a randomized clinical trial to compare a standard of care arm with a new treatment arm for a particular disease. Suppose there are two subpopulations with respect to the response from a specific drug treatment: the responder subgroup and non-responder subgroup. Let m be the number of candidate biomarkers $(Z_1,\ldots,Z_m)$ investigated and n be the total number of patients in the experiments. Let z_{ij} denote the measurement for the j-th candidate biomarker in the i-th patient, and y_i denote the clinical outcome (target variable). The outcome variable can be continuous, binary, or time-to-event onset. Here we will assume the following model for the$j$-th candidate biomarker:

$$h(y_i) = b_{0j} + b_{1j}T + b_{2j}(Tz_{ij}) + \epsilon_{ij}, \quad (1)$$

where T is the treatment indicator and $h(y)$ is a link function. The link function $h(y)$ could be the identity function for continuous outcome, the logistic link for binary response, or the Cox proportional hazard function for time-to-event variables. Here we follow Freidlin and Simon (2005) in omitting the main effect for z_{ij} while retaining only the interaction, but other variations of model formulation will also be considered as part of the project. With Model (1), a significant interaction coefficient b_{2j} indicates a difference in the outcomes between subgroups due to differences in treatment responses in the variable z_{ij}. For candidate biomarkers measured without error, the regression Model (1) is well understood.

However, as discussed in the previous section, for applications like liquid biopsy, it will be common that Z_{ij} is measured with a significant error rate (say from 50% to 98% for sensitivity). Then the question arises to (1) whether the conclusion is still valid regarding the predictive property of Z_{ij} if measurement errors are ignored, and (2) whether it will be helpful to apply a measurement error model to alleviate the problem. A simple answer is not expected. As discussed earlier, the measurement error for deep sequencing has a pretty wide range depending on the technology, patient age, tissue origin, tumor type and stages. Thus, a detailed study over the range of measurement error scenarios will give guidance on when a candidate biomarker has a reasonable chance to be successful, what statistical methods could be used, and what type of applications should be attempted.

Consider the first question, whether the naive approach (simply ignoring the measurement error) can still provide acceptable conclusions. In principle, the naive hypothesis test regarding $b_{2j} = 0$ will still be valid under the null hypothesis for the simple setting considered in Model (1) (see Carroll et al., 2006). However, the power of the test and related sample size considerations could be severely affected. On the other hand, the p-value alone is inadequate to evaluate the candidate biomarker, a good estimate for b_{2j} is a must. Even in simple linear regression, the naive estimate can result in severe bias (the attenuation effect). For link functions other than identity and with other clinical variables, the situation is more complex, including potentially changing the magnitude of the estimate or reversing signs for the regression coefficient. Other than linear models, it can be difficult to derive analytical results for the effect that measurement errors have on the estimate. Thus, Monte Carlo simulation need be used for this purpose.

Now we look at a simple model for survival data to illustrate the approach. Consider proportional hazards model with data from the treatment arm only regarding a single biomarker, the conditional hazard function for survival time Y_i given $\{X_i, Z_i\}$ is

$$\lambda(y|X_i, Z_i) = \lambda_0(y)\exp(\beta_x^T X_i + \beta_z^T Z_i),$$

where X_i is the biomarker state and Z_i represents other variables for the ith patient. We assume X_i to be categorical, which should be applicable to most biomarkers based on deep sequencing in our context. The likelihood is $\ell = \sum_{i=1}^{n} \ell_i$, where

$$\ell_i = \delta_i\{\log\lambda_0(y_i) + \beta_x^T X_i + \beta_z^T Z_i\} - \exp(\beta_x^T X_i + \beta_z^T Z_i) \int_0^{y_i} \lambda_0\,(v)dv.$$

Here X is the true biomarker state, but in reality only the observed state X^* is available through deep sequencing and is subject to error (misclassification). There are two ways to describe the misclassification probability:

$$\pi_{ilk} = P(X^* = x_{(l)}|X_i = x_{(k)}, Z_i) \quad \text{or} \quad \tilde{\pi}_{ilk} = P(X = x_{(l)}|X_i^* = x_{(k)}, Z_i).$$

The range of π or $\tilde{\pi}$ can be inferred from our review in the previous section. If the deep sequencing based biomarker X^* is used without any correction for measurement errors, let

$$\tilde{\ell}_i = \delta_i\{\log\lambda_0(y_i) + \beta_x^T X_i^* + \beta_z^T Z_i\} - \exp(\beta_x^T X_i^* + \beta_z^T Z_i) \int_0^{y_i} \lambda_0\,(v)dv. \tag{2}$$

We have

$$E(\tilde{\ell}_i\,|X^*, Z) = \delta_i\{\log\lambda_0(y_i) + \sum_{k=1}^{r} \tilde{\pi}_{ilk}\,\beta_x^T x_{(k)} + \beta_z^T Z_i\} - \sum_{k=1}^{r} \tilde{\pi}_{ilk} \exp(\beta_x^T x_{(k)} + \beta_z^T Z_i) \int_0^{y_i} \lambda_0\,(v)dv. \tag{3}$$

This can be used to study the bias for the estimate for β as well as the effect on inference. At the same time, a series of measurement error rates covering the range from the SEQC2 and other sources can be used in simulation studies.

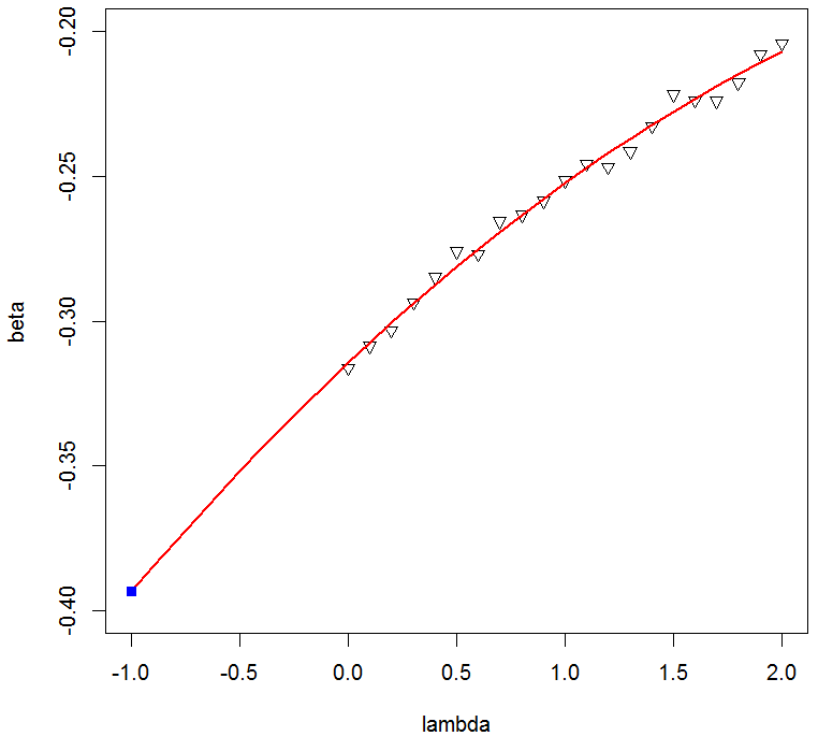

Figure 1. An illustration of the MC-SIMEX approach with a simulated data set. The proportional hazard model is outlined above. In the simulation, data on 300 patients are generated for biomarker positive and negative groups respectively, the true coefficient for Cox regression is -0.4. The estimate for the regression coefficient is calculated for λ values from 0 to 2.0 (triangles in the plot). Note that λ=0 corresponds to the naïve estimator without error correction. The red curve is the resulted quadratic curve fitted to the estimates with different amount of misclassification. The blue square indicates the estimate based extrapolation to λ=-1, which is much closer to the true coefficient value of -0.4.

For the second question, there are multiple approaches. To set the stage, we will continue to use the survival model introduced above and discuss two methods that are of potential interest. The first method is Misclassification Simulation Extrapolation (MC-SIMEX; Küchenhoff et al., 2006). SIMEX method is widely used to handle measurement error problems in various disciplines due to its simplicity and versatility. MC-SIMEX is a variation suited for categorical variables measured with error. With known misclassification rate matrix $\Pi_i = [\pi_{ikl}]_{r\times r}$, we can generate a series of misclassification matrix with larger error rate. Briefly, for a sequence of $\lambda \geq 0$, let $\Pi^{\lambda} = E\Lambda^{\lambda}E^{-1}$, where Λ is the diagonal matrix of eigen values and E is the corresponding matrix of eigen vectors. Then given X^*, we can simulate the biomarker variable with the misclassification rate matrix $\Pi^{1+\lambda}$. The corresponding estimate for regression coefficient $\widehat{\beta}$ can be computed for each λ. With the corresponding $\widehat{\beta}$ and λ values, we can extrapolate to $\lambda = -1$ (corresponding to no measurement error) to obtain the MC-SIMEX estimator for β. Küchenhoff et al. (2006) gives detailed description of the method. The SIMEX method in general has been successfully applied to a variety of problems and is appreciated for its intuitiveness and easiness for implementation. In this case, it is also very attractive in

that it will naturally generate a graphical display for how bias will change with the amount of measurement error or misclassification. This is valuable for practical considerations. As a lower error rate can often be achieved with more expensive technology or by only focusing on late stage tumors, a description of the relationship can be useful for planning subsequent trials.

To illustrate the range of methods that we will consider, we will briefly outline another method, the insertion correction method (Hu et al., 1998; Dupuy, 2005). Follow the notation for our proportional hazards model. Let

$$\ell_i^* = \delta_i\{\log\lambda_0(y_i) + \sum_{k=1}^{r} \pi_{ilk}\,\beta_x^T x_{(k)} + \beta_z^T Z_i\} - \sum_{k=1}^{r} \pi_{ilk}\exp(\beta_x^T x_{(k)} + \beta_z^T Z_i)\int_0^{y_i} \lambda_0\,(v)dv.$$

Then we have $E(\ell_i^*|X_i, Z_i) = \ell_i$. It can be used to derive corrected estimator for β and carry out inference. The large sample distribution can be straightforwardly derived. This necessitates a model for the baseline hazard either parametrically or semiparametrically. It is also possible to work with likelihood or partial likelihood score functions. Both SIMEX and the insertion correction methods can be implemented for binary responses or other settings.

CONCLUSION

Though the NGS technology has tremendous potential in revolutionizing various aspects of precision medicine, its unique characteristics do pose significant challenges for properly analyzing and interpreting the experimental data. In this presentation, we outlined a measurement error based approach for the evaluation of deep sequencing based biomarkers. It provides a reasonable strategy in striking a balance between rigorous regulatory requirements and the need to make effective therapies available to patients in an expeditious manner. Though a huge amount of work need be done to bring this approach in practice, some preliminary results have confirmed the validity of the proposed strategy (Figure 1). More detailed investigation for the utility of this approach will be communicated in the future.

REFERENCES

Bardelli, A., Pantel, K. (2017). Liquid biopsies, what we do not know (yet). Cancer Cell, 31, 172-179.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine, 24, 1713-1723.

Buonaccorsi, J.P. (2010). Measurement Error: Models, Methods, and Applications. Chapman & Hall/CRC.

Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M. (2006). Measurement Error in Nonlinear Models, a Modern Perspective. Chapman & Hall/CRC: Boca Raton.

Chen, J. J., Lu, T. P., Chen, D. T., Wang, S. J. (2014). Biomarker adaptive designs in clinical trials. *Translational Cancer Research*, *3*, 279-292.

Chen, J. J., Lu, T. P., Chen, Y. C., Lin, W. J. (2015). Predictive biomarkers for treatment selection: statistical considerations. *Biomarkers in medicine*, *9*, 1121-1135.

Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., et al. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science, eaar3247.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. Journal of the American Statistical association, 89, 1314-1328.

Deng X., Nakamura Y. (2017). Cancer Precision Medicine: From Cancer Screening to Drug Selection and Personalized Immunotherapy.Trends Pharmacol Sci. 38:15-24.

Druker, B. J., Guilhot, F., O'brien, S. G., Gathmann, I., Kantarjian, H., Gattermann, et al. (2006). Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. New England Journal of Medicine, 355, 2408-2417.

Dupuy, J. F. (2005). The proportional hazards model with covariate measurement error. Journal of statistical planning and inference, 135, 260-275.

FDA. (2016) Optimizing FDA's Regulatory Oversight of Next Generation Sequencing Diagnostic Tests—Preliminary Discussion Paper.
https://www.fda.gov/downloads/MedicalDevices/NewsEvents/WorkshopsConferences/UCM427869.pdf

Foster, J. C., Taylor, J. M., Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. Statistics in medicine, 30(24) 2867-2880.

Freidlin, B., Simon, R. (2005). Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clinical cancer research, 11, 7872-7878.

Gustafson, P. (2004). Measurement Error and Misclassification in Statistics and Epidemiology. Chapman & Hall/CRC, Boca Raton, Florida.

Hoang, M. L., Kinde, I., Tomasetti, C., McMahon, K. W., Rosenquist, T. A., et al. (2016). Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. Proceedings of the National Academy of Sciences, 113, 9846-9851.

Hu, P., Tsiatis, A. A., Davidian, M. (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. Biometrics, 1407-1419.

Hyman, D. M., Taylor, B. S., Baselga, J. (2017). Implementing genome-driven oncology. Cell, 168, 584-599.

Imai K, Ratkovic M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics; 7:443–470.

Krimmel, J. D., Schmitt, M. W., Harrell, M. I., Agnew, K. J., Kennedy, et al. (2016). Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. Proceedings of the National Academy of Sciences, 113, 6005-6010.

Krishnamurthy, N., Spencer, E., Torkamani, A., & Nicholson, L. (2017). Liquid biopsies for cancer: coming to a patient near you. Journal of clinical medicine, 6, 3.

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. Biometrics, 62, 85-96.

Lipkovich I, Dmitrienko A, Denne J, Enas G. (2011) Subgroup identification based on differential effect search (SIDES): a recursive partitioning method for establishing response to treatment in patient subpopulations. Statistics in Medicine; 30:2601–2621.

Oberthuer, A., Juraeva, D., Li, L., Kahlert, Y., Westermann, F., Eils, R., et al. (2010). Comparison of performance of one-color and two-color gene-expression analyses in predicting clinical endpoints of neuroblastoma patients. The pharmacogenomics journal, 10(4), 258.

Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., Posch, M. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *Journal of biopharmaceutical statistics*, *26*, 99-119.

Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M. S., Yeatman, T. J., et al. (2008). Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nature medicine, 14, 822-827.

Sheridan, C. (2017). Grail to pour $1 billion into blood test to detect early cancer. Nature Biotechnology,

35, 101-102.

Shi, L., Kusko, R., Wolfinger, R. D., Haibe-Kains, B., Fischer, M., Sansone, S. A., et al. (2017). The international MAQC Society launches to enhance reproducibility of high-throughput technologies. Nature biotechnology, 35, 1127.

Torga, G., Pienta, K. J. (2017). Patient-Paired Sample Congruence Between 2 Commercial Liquid Biopsy Tests. JAMA Oncology, doi:10.1001/jamaoncol.2017.4027.

Vargas, A. J., Harris, C. C. (2016). Biomarker development in the precision medicine era: lung cancer as a case study. Nature Reviews Cancer, 16, 525-537.

Yang, M., Forbes, M. E., Bitting, R. L., O'Neill, S. S., Chou, P. C., et al. (2017). Incorporating blood-based liquid biopsy information into cancer staging: time for a TNMB system?. Annals of Oncology, 29, 311-323.

Yi, G.Y. (2017). Statistical Analysis with Measurement Error or Misclassification, Strategy, Methods and Application. Springer: New York.

Zhao, Y., Zeng, D., Rush, A. J., Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association, 107, 1106-1118.

CONTACT INFORMATION

Dong Wang
Division of Bioinformatics Biostatistics,
National Center for Toxicological Research
US Food and Drug Administration
3900 NCTR Road
Jefferson, AR 72079
Dong.wang@fda.hhs.gov

Brand and product names are trademarks of their respective companies.